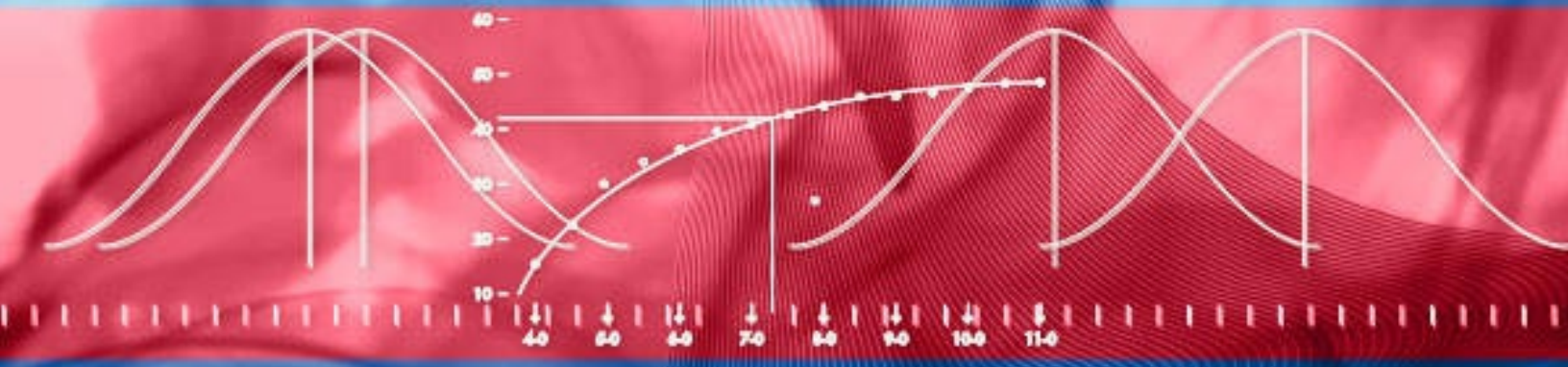
Thomas P. Hogan

Pruebas psicológicas

Una introducción práctica

2ª edición



Manual Moderno®

Pruebas psicológicas
Una introducción práctica

2ª edición

Segunda edición en español traducida de la tercera en inglés

Pruebas psicológicas

Una introducción práctica

Thomas P. Hogan

University of Scranton

Traducción por

Mtro. Jesús Cisneros Herrera

Maestro en Psicología

Universidad Nacional Autónoma de México

Revisión técnica por

Mtra. Laura Edna Aragón Borja

Maestra en Modificación de conducta

Universidad Nacional Autónoma de México

Editor Responsable:

Lic. Santiago Viveros Fuentes

Editorial El Manual Moderno

Manual Moderno®

Editorial El Manual Moderno S.A. de C.V.

Av. Sonora 206 Col. Hipódromo, C.P. 06100 México, D.F.

Editorial El Manual Moderno Colombia S.A.S.

Carrera 12-A No. 79-03/05 Bogotá, DC

Nos interesa su opinión, comuníquese con nosotros:

Editorial El Manual Moderno S.A. de C.V.

Av. Sonora 206, Col. Hipodromo, Deleg. Cuauhtémoc. 06100 México, D.F.

(52-55) 52-65-11-00

info@manualmoderno.com

quejas@manualmoderno.com

Título original de la obra:
Psychological Testing. A Practical Introduction, 3rd edition
Copyright © 2015, 2007, 2003 John Wiley & Sons, Inc.
ISBN: 978-1-118-55412-8 (versión impresa)
"All rights reserved. This translation published under license with the original publisher John & Wiley Sons, Inc."

Pruebas psicológicas. Una introducción práctica, 2ª ed.
D.R. © 2015 por Editorial El Manual Moderno, S.A. de C.V.
ISBN: 978-607-448-498-4 (versión impresa)
ISBN: 978-607-448-501-1 (versión electrónica)

Miembro de la Cámara Nacional de la Industria Editorial Mexicana, Reg. núm. 39

Todos los derechos reservados. Ninguna parte de esta publicación puede ser reproducida, almacenada o transmitida sin permiso previo por escrito de la Editorial.

Para mayor información sobre

Catálogo de producto
Novedades
Distribuciones y más
www.manualmoderno.com

Hogan Thomas P., autor.
Pruebas psicológicas : una introducción práctica / Thomas P. Hogan ; traducción por Jesús Cisneros Herrera. –- 2ª edición. -- México, D.F. : Editorial El Manual Moderno, 2015.
xiv, 497 páginas : ilustraciones ; 28 cm.

Traducción de: Psychological testing : a practical introduction –- 3ª edition.

ISBN: 978-607-448-498-4 (versión impresa)
ISBN: 978-607-448-501-1 (versión electrónica)

1. Pruebas psicológicas – Libros de texto. 2. Psicometría – Libros de texto. I. Cisneros Herrera, Jesús, traductor. II. Título.
150.287-scdd21 Biblioteca Nacional de México

Director editorial y de producción:
Dr. José Luis Morales Saavedra

Editora asociada:
Lic. Vanessa Berenice Torres Rodríguez

Diseño de portada:
LCS Adriana Durán Arce

CONTENIDO

PREFACIO

Propósito y énfasis

Este libro ofrece una introducción al campo de las pruebas psicológicas para el estudiante de psicología y disciplinas afines. Busca ser un método práctico con un énfasis en las estrategias de aprendizaje activo. Su carácter práctico se debe a que aborda las pruebas en su aplicación contemporánea y real en el ejercicio de la psicología. El libro hace hincapié en las estrategias de aprendizaje activo presentando al estudiante los ejercicios ¡Inténtalo! que aparecen a lo largo de todo el texto, en los que se requiere la aplicación de los conceptos y procedimientos que presentamos. Existen demasiados libros de texto sobre pruebas psicológicas que pretenden ser obras de referencia, casi como enciclopedias, más que como verdaderos libros de texto, los cuales deben ser primordialmente un instrumento para el aprendizaje. Las obras de referencia son útiles, pero no como libros de texto, a menos, quizá, que se trate de alumnos avanzados. La investigación sobre el aprendizaje del estudiante ha demostrado de manera amplia que el compromiso activo con el material maximiza el aprendizaje. Hay un uso abundante de las fuentes de internet, pues mucha información que antes era inaccesible a los estudiantes de pruebas psicológicas, y demandaba al profesor esfuerzos sobrehumanos, ahora está disponible en internet. El libro promueve el uso de estos recursos. Además de los ejercicios incorporados directamente en el texto, cada capítulo empieza con una lista de objetivos de aprendizaje y concluye con un resumen de los puntos importantes, una lista de palabras clave y ejercicios adicionales. También incluimos resúmenes de puntos clave para reforzar el material importante dentro de los capítulos. Estos resúmenes intermedios deben ayudar al estudiante a organizar la información. Todas estas características deben ayudar al aprendizaje del estudiante.

La lista de objetivos al principio de cada capítulo debe servir como "organizador avanzado" para ayudar al estudiante a enfocar su atención. El resumen al final de cada capítulo ayudará a hacer el cierre, a veces después de una travesía difícil a lo largo del material. La lista de palabras clave debe complementar los puntos del resumen. Al final de cada capítulo se agregan abundantes ejercicios cuyo carácter varía: algunos hacen hincapié en cuestiones técnicas, otros en preguntas "pensadas" y otras más piden al estudiante encontrar información en los recursos de internet. No esperamos que el estudiante realice todos los ejercicios; sin embargo, el profesor puede hacer una selección razonable de ejercicios que el estudiante realice de manera individual o grupal. Los ejercicios al final de los capítulos son más desafiantes que los que aparecen en las secciones ¡Inténtalo! que aparecen en el texto principal, pero traté de diseñarlos, en su

mayoría, de modo que puedan terminarse en 10 o 20 minutos, algunos incluso en 2 o 3. Los estudiantes encontrarán que realizar, al menos, algunos de estos ejercicios les ayuda a comprender y retener el material.

El acento en la utilidad práctica del libro no implica una falta de rigor o la evitación de temas difíciles; por el contrario, el texto aborda el material difícil de manera frontal. El uso de pruebas psicológicas no es un tema fácil y el libro no pretende empobrecer intelectualmente el material. Sin embargo, si busca mostrar la práctica real en el campo de las pruebas psicológicas y dar ejemplos de conceptos y procedimientos empleados en las aplicaciones contemporáneas.

Formación del estudiante

Mientras preparaba el texto, di por hecho que el estudiante habría terminado un curso de estadística básica en el que se incluyeran los siguientes temas: métodos para resumir datos crudos en cuadros y gráficas, formas de las distribuciones, medidas de tendencia central y de variabilidad, correlación (de orden cero) y regresión, elementos de la teoría del muestreo, estimación de intervalos y prueba de hipótesis. También di por hecho que el estudiante habría olvidado una cantidad considerable de este material, por lo que el texto presenta "repasos" de varios temas de estadística básica a lo largo de los primeros capítulos. Debido a que los estudiantes tienen la predisposición a evitar cualquier cosa que aparezca en un apéndice, incorporé estos repasos en el texto principal. El estudiante que, en realidad, recuerde mucho de lo que se trata de estadística básica, que es muy poco común encontrar, puede omitir los repasos. Los profesores necesitarán usar su criterio acerca de cuánto tiempo dedicar a los repasos. Además, el texto aborda ciertos temas de estadística multivariada, en especial la correlación múltiple y el análisis factorial. Sin embargo, la mayoría de los estudiantes no habrán revisado estos temas, por lo que no di por hecho ningún conocimiento previo.

Organización

Este libro se divide, por su propia naturaleza, en dos secciones principales. La primera parte trata los conceptos básicos del campo de las pruebas psicológicas. En el capítulo 1, se presenta un panorama general del campo, incluyendo los usos típicos de las pruebas psicológicas. El capítulo 2 constituye un tratamiento mucho más completo de las fuentes de información acerca de las pruebas del que podría encontrarse en cualquier otro lugar. He hecho esto por dos razones; primero, los estudiantes, así como otros, con frecuencia hacen la pregunta "¿Hay una prueba que mida...?" Y yo quiero que los estudiantes que usen este texto sepan cómo responder esta pregunta común. Segundo, en capítulos posteriores, se pide a los estudiantes que usen las fuentes de información citadas en el capítulo 2 para encontrar ejemplos de la vida real de conceptos, procedimientos y pruebas.

En los capítulos 3 a 6 se revisan los temas fundamentales de normas, confiabilidad, validez, elaboración de pruebas y neutralidad. En cada uno de estos capítulos, la mayor parte del material debe revisarse con todos los estudiantes, pues el resto el material es un poco más avanzado o más técnico. Los profesores pueden incluir todo este material, parte de él o nada, dependiendo de sus áreas de interés y las necesidades específicas de sus alumnos. El capitulo 6, que trata de la elaboración de pruebas, se complementa con el apéndice B con los pasos para construir una prueba sencilla. Algunos profesores piden a sus estudiantes construir una prueba sencilla como ejercicio de clase, y el apéndice B debe ser de utilidad en este caso. Los pequeños conjuntos de datos del apéndice D pueden usarse con paquetes estadísticos estándar para hacer ejercicios prácticos relacionados con los conceptos de los capítulos 3-6. Ejercicios selectos al final de los capítulos requieren usar estos datos, pero algunos profesores preferirán usar sus propios datos.

El equilibrio que deseamos alcanzar entre la teoría clásica de las pruebas (TCP) y la teoría de la respuesta al reactivo (TRR) en los capítulos 3-6 constituye un desafío especial. La TCP es lo suficientemente difícil para los novatos, mientras que la TRR resulta asfixiante. En la práctica actual, los procedimientos de la TRR se aplican de manera rutinaria. Los estudiantes encontrarán los procedimientos de la TRR y la TCP en casi todas las pruebas elaboradas en años recientes. Incluso el principiante, necesita conocer los conceptos y el lenguaje de la TRR, por lo que, de acuerdo con mi intención de ser práctico, los capítulos 3-6 contienen una buena dosis de procedimientos de la TRR y la TCP. Desde luego, los profesores necesitarán lograr el equilibrio correcto de acuerdo con las necesidades de sus alumnos.

La segunda parte ofrece una introducción a las principales categorías de las pruebas psicológicas. Un capítulo bosqueja los principales métodos conceptuales y de procedimientos de cada categoría y presenta algunos ejemplos de pruebas. He tratado de resistirme a la tentación de enumerar prueba tras prueba con breves descripciones de cada una, porque no creo que los estudiantes aprendan algo útil de un catálogo de esa naturaleza. De hecho, tal vez no aprendan nada de tales listas. Por lo común, cuando una prueba aparece en la segunda parte, he tratado de describirla con suficientes detalles, de modo que el estudiante principiante pueda aprenda algo de ella. Hice sólo algunas excepciones a esta regla cuando se trató de casos extraordinarios. La selección de ejemplos para los capítulos de la segunda parte estuvo guiada *primordialmente* por la frecuencia del uso real de las pruebas –de acuerdo con la intención de ofrecer un libro práctico– y de manera secundaria, por el deseo de ilustrar cierta diversidad de métodos en una categoría de pruebas. Cuando revisen los capítulos de la segunda parte, espero que los profesores cuenten con los materiales de las pruebas para poder mostrarlos a sus alumnos. Sin embargo, si esto no es posible, pero cuentan con materiales de otras pruebas, quizá sea preferible usar estas otras pruebas en vez de las que aparecen en el texto. Es especial, tratándose de novatos en el campo, no es una experiencia significativa leer sobre una prueba sin verla ni sentirla literalmente.

La segunda parte concluye con un capítulo, el 16, dedicado a los aspectos éticos y

legales. Este capítulo no se ajusta temáticamente a los demás de la segunda parte, pero es evidente que es necesario. Pienso que los estudiantes no pueden apreciar algunos de los temas tratados en este capítulo sino después de haber revisado todos los demás capítulos. No quería una tercera parte de sólo un capítulo, que sería como el prohibido párrafo de una sola línea, así que incluí el capítulo sobre aspectos éticos y legales al final de la segunda parte.

Lo nuevo de la tercera edición

La tercera edición preserva las principales características de la segunda edición, pero introduce los siguientes cambios importantes:

- **Expansión considerable del tratamiento de la neutralidad/sesgo de las pruebas:** La neutralidad de las pruebas ha ascendido de manera constante en la constelación psicométrica de los temas esenciales. Las primeras ediciones del libro cubrían el tema, pero en secciones dispersas que ahora se han reunido, coordinado y expandido. La neutralidad, de manera lógica, pertenece a la validez; sin embargo, como se señala en el texto, gran parte del trabajo práctico relacionado con la neutralidad ocurre durante el proceso de elaboración de la prueba, por lo que decidí incluir el material sobre este tema en el capítulo 6, dedicado a la elaboración de pruebas.
- **Coordinación con el nuevo *Standards for Educational and Pschological Testing*:** El nuevo *Standards*, que estuvo en revisión durante casi cinco años, apareció casi al mismo tiempo que esta nueva edición que, de hecho, buscó de manera intencional reflejarlo. A lo largo del texto aparecen citas del nuevo *Standards*, en especial en los capítulos 3-6.
- **Actualización minuciosa de las nuevas versiones de pruebas muy usadas:** Teniendo en cuenta el vertiginoso ritmo actual de la elaboración de pruebas, mantenerse al día fue quizá la parte más abrumadora al preparar esta nueva edición. Más o menos la mitad de todas las pruebas que aparecen en esta tercera edición son pruebas nuevas o revisiones recientes. Entre ellas, se encuentran las siguientes (por ahora nos limitamos a presentar las iniciales):

  WAIS-IV
  WMS-IV
  PPVT-4
  WPT
  GRE
  ASVAB
  NBAP-D
  MMSE-2

NEO PI-3
MMPI-2 RF

El apéndice C incluye información actualizada de contacto de las principales editoriales de pruebas. Prácticamente todo contacto con las editoriales empieza mediante internet, por lo que se incluyen las URL de las editoriales y se omiten direcciones y números telefónicos.

Además de los conjuntos de datos muestra que aparecen en el apéndice D, se incluye una hoja de cálculo de Excel que permite al estudiante generar curvas características de reactivo (CCR) variando los parámetros de los reactivos en el modelo de tres parámetros.

El manual del profesor y el banco de pruebas se han actualizado con meticulosidad de acuerdo con los cambios en esta nueva edición.

Al estudiante

Si vas a trabajar en algún área de la psicología, casi sin duda tendrás que tratar con pruebas psicológicas, por lo que aprender acerca de ellas tiene aplicaciones prácticas. Espero que este texto te ayude a comprender los temas básicos del campo de las pruebas psicológicas.

He aquí mis sugerencias para usar el texto con eficacia.

- Revisa los objetivos al principio de cada capítulo para estar alerta al material importante.
- Observa con cuidado las "palabras clave"; se denominan así porque son esenciales. Aparecen en negritas en el texto y se encuentran en la lista al final de cada capítulo y en el glosario al final del libro.
- Realiza todos los ejercicios ¡Inténtalo! esparcidos a lo largo del texto. Hacerlo ayudará a que el material "penetre". Observa que cada ejercicio toma más o menos un minuto.
- Usa los resúmenes de puntos clave intermedios que aparecen a lo largo del texto para ayudar a que enfoques tu atención y organices el material. Cuando llega a uno de ellos, haz una pausa para hacer un repaso del material que le precede.
- Usa los puntos de los resúmenes al final de cada capítulo para repasar los principales temas y puntos que se revisan en él.
- Realiza, al menos, algunos ejercicios al final de cada capítulo. No se espera que los realices todos, pero haz algunos. Al igual que los ejercicios ¡Inténtalo!, estos ayudarán a hacer práctico el material.
- Por último, observa que las pruebas psicológicas no son un tema fácil. ¡Estudia con tenacidad!

Reconocimientos

Dar cuenta de la gran cantidad de contribuciones para la preparación de este libro es una tarea abrumadora y un acto de humildad, pues numerosas personas han hecho mucho para ayudarme. Estoy muy agradecido, en especial con las siguientes personas. Con todos mis estudiantes a lo largo de muchos años por su buena disposición para sugerir maneras eficaces de presentar conceptos del campo de las pruebas; mi especial agradecimiento a Allyson Kiss y Matthew Sabia por su ayuda en la investigación y la preparación del manuscrito de esta tercera edición. Con muchas editoriales por concederme la autorización para reproducir su material y a sus equipos de trabajo que me dieron útiles consejos acerca de sus pruebas. Agradezco en especial a los siguientes individuos que me ofrecieron retroalimentación sobre su uso real de los libros, así como comentarios sobre el plan de revisiones de ésta y las ediciones anteriores: Ira h. Bernstein, University of Texas, Arlington; Jeffrey B. Brookings, Wittenberg University; Douglas Maynard, SUNY, New Paltz; Robert Resnick, Randolph Macon College; and Marie D. Thomas, California State University, San Marcos. Además, mis agradecimientos a los siguientes individuos, que hicieron comentarios y sugerencias útiles sobre los capítulos revisados de ésta y las ediciones anteriores: Julie Alvarez, Tulane University; David Bush, Utah State University; Mark Lenzenweger, State University of New York, Binghamton; Stuart McKelvie, Bishop's University; John Suler, Rider University; Stefan Schulenberg, University of Mississippi; y David Trumpower, Marshall University. Todos ellos ayudaron a crear un mejor libro de texto.

En el largo plazo, mi permanente gratitud es para mi mentora académica, la renombrada Anne Anastasi; para Dorothea McCarthy, que hizo los arreglos para mi primer trabajo en el campo de las pruebas, y para Joseph Kubis por su don para la pedagogía. Para mis mentores profesionales, Dena Wadell y Roger Lennon, quienes me mostraron la conexión entre teoría y práctica. Para mis colegas, empezando con William Tsushima, cuya ayuda al inicio de mi carrera fue más importante de lo que él puede imaginar. Agradecimientos muy especiales para mi colega de la University of Scranton, John Norcross, que fungió como caja de resonancia de una gran cantidad de temas. Y desde luego, para Brooke Cannon y Matthew Eisenhard por su excelente contribución con el capítulo 10 sobre evaluación neuropsicológica. Por último, quiero agradecer a mi esposa, Peg, y a nuestros hijos por su apoyo moral (y algunas sugerencias en verdad útiles) a lo largo de esta empresa.

Con toda esa ayuda, podría pensarse que el libro es perfecto. Desafortunadamente, tal vez no sea así. Así que asumo la responsabilidad por cualquier imperfección que se haya filtrado en este trabajo.

Thomas P. Hogan
Mayo de 2013
Scranton, Pennsylvania

PRIMERA PARTE

La primera parte de este libro ofrece un panorama general del campo de las pruebas psicológicas y, después, se enfoca en los principios y procedimientos fundamentales que se pueden aplicar a todo tipo de pruebas. El capítulo 1 introduce al estudiante en el campo actual y presenta un esbozo de cómo se llegó a ese estado. En el capítulo 2 se revisan las fuentes de información a las que el estudiante puede recurrir para averiguar más sobre las pruebas. Dichas fuentes se usan con frecuencia en los capítulos posteriores para identificar qué pruebas se emplean con propósitos específicos; por ello, es importante aprender a usarlas.

En los capítulos 3 al 6 se presentan los principios fundamentales para juzgar cualquier prueba. Estos capítulos se ocupan de las normas (3), confiabilidad (4), validez (5) y elaboración de pruebas, incluyendo la neutralidad (6). Aunque no se trata de un material sencillo, es esencial que el estudiante aprenda estos conceptos básicos, ya que son la base para manejar las pruebas que se tratan en la segunda parte de este libro. Para asimilar el material, el estudiante debe hacer los ejercicios INTÉNTALO, que aparecen a lo largo del texto. Los capítulos 1 y 2 pueden leerse de manera relajada, pero los capítulos 3 al 6 requieren de mucha concentración y análisis. Estudiar estos capítulos apropiadamente tendrá una recompensa, pues dará buenos dividendos al leer los capítulos posteriores.

CAPÍTULO 1
Campo de las pruebas psicológicas

Objetivos

1. Enumerar las principales categorías de pruebas presentando, al menos, un ejemplo de cada una.

2. Identificar los principales usos y usuarios de las pruebas.

3. Resumir los principales supuestos y las preguntas fundamentales relacionadas con las pruebas.

4. Bosquejar las características importantes de los principales períodos de la historia de las pruebas.

5. Identificar los seis principales acontecimientos que influyeron en el desarrollo de las pruebas.

6. Dar una definición de "prueba".

Introducción

Este capítulo ofrece un panorama general del campo de las pruebas. Desde luego, todos saben, al menos en general, qué significa "prueba". Todos tienen alguna experiencia con diversas pruebas, por ejemplo, pruebas de admisión a la universidad, exámenes finales de cursos, inventarios de intereses vocacionales y, quizá, algunas medidas de personalidad. Sin embargo, al emprender el estudio formal de este campo, es importante adquirir una comprensión más amplia y precisa de él. "Más amplia" implica considerar todos los tipos de pruebas y todos los temas pertinentes: no queremos omitir nada importante. "Más precisa" significa adquirir el dominio técnico indispensable para los profesionales de la extensa área de la psicología y disciplinas afines: no estaremos satisfechos con sólo dar a conocer estos temas.

Ésta es una agenda ambiciosa para un capítulo; sin embargo, en este primer capítulo sólo intentamos brindar un panorama general de estos temas, mientras que en los restantes proporcionaremos detalles. Hay distintos modos de cumplir con nuestro objetivo, pero ninguno es el mejor, por lo que recurrimos a cinco perspectivas o aproximaciones para introducirnos al campo de las pruebas, es decir, lo examinamos desde diferentes ángulos o con distintos lentes. Primero, bosquejamos las principales categorías de pruebas, las cuales, en su mayoría, corresponden a capítulos de la segunda parte del libro. Al describir estas categorías, presentamos ejemplos de algunas de las pruebas más empleadas. En segundo lugar, identificamos los principales usos y usuarios de las pruebas. ¿Quién las usa y con qué propósitos? En tercer lugar, bosquejamos los asuntos primordiales que nos preocupan en relación con las pruebas. Este bosquejo, es decir, la lista de las principales preocupaciones, corresponde a los capítulos de la primera mitad del libro. En cuarto lugar, rastreamos las raíces históricas del estado actual del campo de las pruebas. Distinguimos los principales períodos de esta historia e identificamos algunas de las principales fortalezas que han moldeado este campo. Por último, examinamos algunos de los intentos de definir "prueba" y otros términos relacionados. Después de haber revisado el campo de las pruebas desde estas cinco perspectivas, será posible tener un panorama general.

Resumen de puntos clave 1-1

Cinco maneras de introducirnos al campo

1. Categorías de las pruebas
2. Usos y usuarios de las pruebas
3. Supuestos y preguntas fundamentales
4. Períodos históricos y fortalezas
5. Definición

Principales categorías de las pruebas

Empezamos nuestra exploración del campo de las pruebas identificando las principales categorías en que se agrupan. Cualquier clasificación de este tipo necesariamente tiene límites difusos, pues las categorías a menudo se mezclan entre sí en vez de diferenciarse con claridad. No obstante, un esquema organizacional nos ayuda a comprender la amplitud del campo. El Resumen de puntos clave 1-2 ofrece el esquema de clasificación que usamos en todo el libro; de hecho, los capítulos 8 al 15 siguen esta organización. Este capítulo introductorio sólo toca las principales categorías, pero cada una de ellas será tratada en profundidad más adelante.

La primera categoría abarca las **pruebas de capacidad mental**. En el campo de las pruebas psicológicas, el término "capacidad mental" incluye numerosas funciones cognitivas, como memoria, visualización espacial o pensamiento creativo. A lo largo de la historia, esta área se ha centrado en la inteligencia, definida ampliamente. Esta categoría se subdivide en pruebas de inteligencia de aplicación individual, pruebas de inteligencia de aplicación grupal y otras pruebas de capacidades distintas a las de inteligencia. Un ejemplo de las pruebas de inteligencia de aplicación individual es la Escala Wechsler de Inteligencia para Adultos[1], WAIS por sus siglas en inglés. Otro ejemplo clásico de esta categoría es la Escala de Inteligencia Stanford-Binet. Estas pruebas son administradas por psicólogos bien capacitados para realizar evaluaciones individuales, esto es, de uno a uno, con el objetivo de proporcionar un índice de la capacidad general mental de un individuo. Un ejemplo de una prueba de inteligencia de aplicación grupal es el *Otis-Lennon School Ability Test* (OLSAT [Prueba de Capacidad Escolar Otis-Lennon]), la cual se aplica a un grupo de estudiantes, por lo general en un salón de clases, para estimar la capacidad mental para tener un buen desempeño en asignaturas escolares típicas. Otra prueba de esta categoría es el **SAT**[2], que se usa para predecir el éxito en la universidad.

¡Inténtalo!

Para profundizar más en una categoría, pasa a la página 291a». Echa un vistazo rápido a las páginas 291-300a». Verás cómo los capítulos subsiguientes brindan detalles acerca de las pruebas mencionadas en este capítulo.

Hay muchos otros tipos de pruebas de capacidad mental –casi podríamos decir infinitos–, como las de memoria, razonamiento cuantitativo, pensamiento creativo, vocabulario y capacidad espacial. A veces, estas funciones mentales se incluyen en pruebas de capacidad mental general; otras veces, constituyen por sí mismas pruebas específicas para medir tales capacidades de manera independiente.

La siguiente categoría principal abarca las **pruebas de rendimiento**, las cuales intentan evaluar el nivel de conocimiento o habilidad de un individuo en un dominio específico.

Aquí sólo tratamos las pruebas elaboradas de manera profesional y estandarizadas, y excluimos una amplia serie de pruebas hechas por maestros para usarlas de manera cotidiana en su labor educativa. Incluso excluyendo estas últimas, las pruebas de rendimiento se emplean con facilidad y son las más usadas. La primera subdivisión de esta categoría incluye las baterías utilizadas en las escuelas primarias y secundarias. Entre éstas se encuentran el *Stanford Achievement Tes*t [Prueba de Rendimiento Stanford], el *Metropolitan Achievement Tests* [Prueba de Rendimiento Metropolitan] y el *Iowa Tests of Basic Skills* [Prueba Iowa de Habilidades básicas], las cuales constan de una serie de pruebas en áreas como lectura, matemáticas, lenguaje, ciencia y ciencias sociales. La segunda subdivisión incluye pruebas de un solo tema, es decir, explora una sola área, como psicología, francés o geometría.

Un ejemplo de tales pruebas es la Prueba Psicológica: *Graduate Record Examinations* (GRE [Exámenes de Registro para Graduados]).

Resumen de puntos clave 1-2

Principales categorías de las pruebas

Pruebas de capacidad mental
- De aplicación individual
- De aplicación grupal
- Otras capacidades

Pruebas de rendimiento
- Baterías
- Tema único
- Certificación, licencia
- Programas con financiamiento gubernamental
- Pruebas de rendimiento individual

Pruebas de personalidad
- Pruebas objetivas
- Técnicas proyectivas
- Otros enfoques

Intereses y actitudes
- Intereses vocacionales
- Escalas de actitud

Pruebas neuropsicológicas

La tercera subdivisión incluye las innumerables pruebas que se usan para la certificación y concesión de licencias para ejercer una profesión, como enfermería, enseñanza, terapia física o ser piloto de líneas comerciales. Ninguna de estas pruebas es conocida más allá de su ámbito, pero tienen consecuencias importantes en los campos vocacionales específicos.

En cuarto lugar, varias agencias gubernamentales financian ciertos programas de pruebas de rendimiento, entre las cuales los programas estatales en temas básicos como lectura, escritura y matemáticas son los más notables. De hecho, estos programas de

evaluación estatal han cobrado gran importancia en años recientes como consecuencia de las nuevas leyes federales. En en algunos estados de EUA, graduarse del bachillerato depende, en parte, del desempeño en estas pruebas. Otros programas con financiamiento gubernamental ofrecen información acerca del desempeño nacional en distintas áreas. Los intentos más conocidos son el *National Assessment of Educational Progress* (NAEP [Evaluación Nacional de Progreso Educativo]) y el *Trends in International Mathematics and Science Study* (TIMSS [Tendencias en el Estudio Internacional en Matemáticas y Ciencia]), sobre los cuales se informa con frecuencia en los medios.

Por último, hay pruebas de rendimiento que se aplican de manera individual. Los primeros cuatro tipos de las pruebas de rendimiento se aplican, por lo general, en grupo; sin embargo, algunas de ellas se aplican de modo individual de manera muy parecida a la de las pruebas de capacidad mental. Las pruebas de aplicación individual ayudan en el diagnóstico de trastornos como los problemas de aprendizaje.

La siguiente categoría principal abarca diferentes pruebas diseñadas para obtener información sobre la personalidad humana. La primera subdivisión incluye las llamadas **pruebas objetivas de personalidad**. En el lenguaje de este campo, objetivo significa únicamente que la calificación es objetiva, es decir, está basada en reactivos que se responden con verdadero o falso o en un formato similar. Ejemplos de estas pruebas son el Inventario Multifásico de Personalidad de Minnesota (MMPI por sus siglas en inglés), el *Beck Depression Inventory* (BDI [Inventario de Depresión de Beck]) y el *Eating Disorder Inventory* (EDI [Inventario de Trastornos Alimenticios]). El MMPI ofrece un perfil que muestra qué tan similares son las respuestas del examinado a las de distintos grupos clínicos. El BDI y el EDI, como lo indican sus nombres, miden depresión y trastornos de la alimentación, respectivamente. Por comodidad y claridad conceptual, en los subsiguientes capítulos dividimos estas pruebas objetivas en las que miden rasgos de personalidad dentro del rango normal y las que se diseñaron como instrumentos clínicos para medir los padecimientos patológicos e incapacitantes.

¡Inténtalo!

Parte de convertirse en un profesional de este campo implica aprender las iniciales de las pruebas, pues aparecen de manera rutinaria en informes psicológicos y artículos de revista, a menudo, sin referencia a su nombre completo. ¡Acostúmbrate a esto! Sin ver otra vez el texto, trata de recordar a qué prueba corresponden las siguientes iniciales:

EDI _____
GRE _____
WAIS _____
MMPI _____

La segunda subdivisión principal de las pruebas de personalidad incluye las **técnicas proyectivas**, que consisten en una tarea relativamente simple o no estructurada. Se espera que las respuestas del examinado revelen algo acerca de su personalidad. La más

famosa de estas técnicas es el Prueba Rorschach de Manchas de Tinta, a veces sólo llamada Rorschach o prueba de manchas de tinta. Otros ejemplos son las técnicas de dibujos de la figura humana, frases incompletas o reacciones ante imágenes. Incluimos en las medidas de personalidad una tercera categoría, que rotulamos simplemente como "otros enfoques", para englobar la miríada de modos que los psicólogos han concebido para satisfacer nuestra inagotable fascinación por la personalidad humana.

La siguiente categoría principal de pruebas abarca las medidas de intereses y actitudes, cuya subdivisión más notable es la de **medidas de intereses vocacionales**. Estas pruebas son ampliamente usadas en el bachillerato y en las universidades para ayudar a los estudiantes a explorar los trabajos acordes con sus intereses. Ejemplos de estas pruebas son el *Strong Interest Inventory* (SII [Inventario de Intereses Sólidos]) y el Kuder Career Search (KCS [Búsqueda de Carrera Kuder]). Esta categoría también incluye numerosas medidas de actitudes hacia temas, grupos y prácticas; por ejemplo, hay medidas para la actitud hacia la pena capital o hacia los ancianos.

La última categoría está conformada por las **pruebas neuropsicológicas**, cuyo propósito es ofrecer información acerca del funcionamiento del sistema nervioso central, en especial del cerebro. Desde algunas perspectivas, ésta no debería ser una categoría separada porque muchas pruebas empleadas con este propósito vienen de otras categorías, como las pruebas de capacidad y de personalidad. Sin embargo, empleamos una categoría separada para agrupar pruebas usadas de manera específica para evaluar las funciones cerebrales. Las pruebas de memoria para el material verbal y visual, de coordinación psicomotriz y de pensamiento abstracto son de especial interés.

¡Inténtalo!

Aquí hay una prueba sencilla que usan los neuropsicólogos; se llama cruz griega. Mira la figura por un momento, luego, cúbrela y trata de dibujarla de memoria. ¿Qué conductas y procesos mentales crees que están involucrados en esta prueba?

Otras maneras de clasificar las pruebas

Hasta aquí, hemos clasificado las pruebas de acuerdo con el tipo predominante de contenido; de hecho, ésta es la manera más común y, para la mayoría de perspectivas, más útil de hacerlo. Sin embargo, hay otras, las cuales enumeraremos brevemente. Se pueden ver en el Resumen de puntos clave 1-3.

Resumen de puntos clave 1-3

Otras maneras de clasificar las pruebas

- Lápiz y papel versus ejecución
- Velocidad versus poder
- Pruebas individuales versus pruebas grupales
- Ejecución máxima versus ejecución típica
- Referidas a la norma versus referidas al criterio

Pruebas de lápiz y papel versus pruebas de ejecución

En una **prueba de ejecución**, el examinado realiza alguna acción como armar un objeto, producir un discurso, llevar a cabo un experimento o guiar a un grupo, mientras que en una **prueba de lápiz y papel**, responde a un conjunto de preguntas, por lo general, usando papel y lápiz, como lo dice su nombre. Muchas de estas pruebas tienen formatos de respuesta de opción múltiple, de falso-verdadero u otros semejantes; en la actualidad, frecuentemente se pueden realizar en una computadora haciendo clic para responder.

Pruebas de velocidad versus pruebas de poder «7

El propósito fundamental de las **pruebas de velocidad** es ver qué tan rápido se desempeña el examinado. La tarea, por lo general, es sencilla y la puntuación depende de la cantidad de reactivos o tareas terminados en un tiempo límite o de cuánto tiempo (p. ej., minutos o segundos) requiere el examinado para realizar la tarea. Por ejemplo, ¿cuánto tiempo tardarías en tachar todas las letras "e" de esta página? ¿Cuánto tardarías en resolver 50 problemas sencillos de aritmética, como 42 + 19 o 24 × 8? Por otra parte, una **prueba de poder**, por lo general, implica material desafiante y no hay límite de tiempo o éste es muy generoso. El punto esencial de estas pruebas es evaluar los límites del conocimiento o capacidad del examinado (no importa la velocidad). La distinción no necesariamente es tan tajante: velocidad pura o poder puro. Algunas pruebas de poder pueden tener un elemento de velocidad; no se puede contestar el SAT eternamente; sin embargo, la capacidad mental y el conocimiento, más que la velocidad, son los determinantes primordiales del desempeño en una prueba de poder. Algunas pruebas de velocidad pueden tener un elemento de poder; se tiene que pensar un poco y, quizá, también hacer un plan para tachar todas las "e" de esta página. Sin embargo, esta tarea es primordialmente cuestión de velocidad, no de conocimiento científico.

Pruebas individuales versus pruebas grupales

Esta distinción se refiere simplemente al modo en que se aplica la prueba. Una **prueba**

individual puede aplicarse sólo a un examinado a la vez. Los ejemplos clásicos son las pruebas de inteligencia en las que un examinador presenta cada pregunta o tarea al individuo y registra sus respuestas. Una **prueba grupal** se puede aplicar a varios individuos al mismo tiempo, es decir, a un grupo; desde luego, cada uno de ellos recibe su propia puntuación. En general, cualquier prueba de aplicación grupal puede aplicarse a un solo individuo a la vez cuando las circunstancias lo ameritan, pero una de aplicación individual nunca se puede aplicar a un grupo a la vez.

Ejecución máxima versus ejecución típica

Ésta es otra distinción útil entre tipos de pruebas. Algunas buscan la **ejecución máxima**: ¿qué tan bueno es el desempeño del examinado cuando hace lo mejor que puede? Por lo general, esto sucede con las pruebas de rendimiento y capacidad. Por otro lado, a veces queremos ver la **ejecución típica** del individuo, como en las pruebas de personalidad, intereses y actitudes. Por ejemplo, en una prueba de personalidad queremos saber qué tan extrovertido es regularmente el examinado, no cuán extrovertido puede ser si en verdad se esfuerza por parecer extrovertido.

Referidas a la norma versus referidas al criterio

Muchas pruebas tienen normas basadas en el desempeño de los casos en un programa de estandarización. Por ejemplo, si la puntuación de un individuo en el SAT o ACT está en el percentil 84, significa que su puntuación es mejor que la de 84% del grupo nacional del que se obtuvieron las normas. Ésta constituye una **interpretación referida a la norma** del desempeño en una prueba. En contraste, algunas interpretaciones dependen de algún criterio definido con claridad que se usa como referencia y no de un conjunto de normas; por ejemplo, un instructor puede decir: quiero que se aprendan todos los términos clave que están al final del capítulo. Si en la prueba del instructor un alumno define correctamente sólo 60% de los términos, se considera insuficiente sin importar los resultados del resto de los alumnos. Ésta es una **interpretación referida al criterio.** En realidad, es el método de interpretación más que la prueba en sí misma lo que se puede calificar como referida a la norma o referida al criterio. Exploramos esta distinción con todo detalle en el capítulo 3.

Usos y usuarios de las pruebas

Un segundo modo de introducirse en el campo de las pruebas es identificar los usos y usuarios regulares de las pruebas. ¿Quién usa las pruebas ubicadas en las categorías enumeradas en la sección anterior? ¿En qué escenarios? Consideremos los siguientes ejemplos.

- John es psicólogo clínico que ejerce en el ámbito privado, donde ve muchos clientes que sufren de ansiedad y depresión. Algunos casos pueden ser de intensidad moderada susceptibles de terapia conductual y cognitivo conductual de corto plazo, pero otros pueden ser más crónicos con síntomas que encubren un padecimiento potencialmente esquizofrénico. Al principio de la evaluación de sus clientes, John emplea de manera rutinaria el MMPI y, en casos muy complicados, la Prueba Rorschach de Manchas de Tinta.
- Kristen es psicóloga educativa. Cuando un maestro le envía algún alumno, por lo general revisa los registros escolares que incluyen las puntuaciones del *Otis-Lennon School Ability Test* y el *Stanford Achievement Test*. Además, aplica la Escala Wechsler de Inteligencia para Niños (WISC) y alguna escala para evaluar la conducta.
- Frank es consejero de bachillerato y supervisa la aplicación anual del Strong Interest Inventory (SII) por parte de la escuela. Los resultados de la prueba se reparten en los salones donde se pasa lista a los alumnos. Ya que Frank no se puede reunir con cada uno para entregarle sus resultados, prepara materiales para los maestros que pasan lista para que puedan ayudar a los alumnos a interpretar sus informes.
- Annika es psicóloga del desarrollo que se interesa en el estrés infantil que se presenta cuando los chicos y chicas pasan de la etapa prepuberal a la adolescencia. En su estudio longitudinal, usa una medida del autoconcepto (*Piers-Harris Children's Self-Concept Scale* [Escala Piers-Harris de Autoconcepto Infantil]) para rastrear los cambios en cómo se sienten los participantes consigo mismos. También cuenta con puntuaciones de pruebas de inteligencia que toma de los registros escolares sólo para describir la naturaleza de su muestra.
- Brooke es neuropsicólogo. En una demanda legal contra un productor de automóviles, un individuo declaró haber sufrido daño cerebral en un accidente. En defensa del productor, Brooke presentó evidencia, obtenida de distintas pruebas, de que no había tal daño cerebral.
- Bill es asistente del director de recursos humanos en la compañía MicroHard, la cual contrata casi 100 nuevas secretarias al año en sus diferentes sucursales. Bill supervisa la aplicación de pruebas a 500 candidatas cada año y trata de asegurarse de que tienen las habilidades tanto técnicas como interpersonales que les permitirán ser miembros productivos del "equipo MicroHard".
- Joe trabaja para el Departamento Estatal de Educación. La asamblea legislativa adoptó apenas un proyecto de ley en el que se exige que todos los estudiantes aprueben exámenes de lectura, matemáticas y escritura para poder recibir su certificado de bachillerato. Joe –un psicólogo afortunado–, tiene que organizar la preparación de estas pruebas.

Todos éstos son ejemplos de los usos y usuarios típicos de las pruebas, pero vamos a

presentar un catálogo más sistemático. Como se enumera en Resumen de puntos clave 1-4, se identifican cuatro grupos principales de usuarios; aunque hay una diversidad considerable en cada grupo, cada uno es diferente en la manera en que emplea las pruebas. También notamos que cada grupo usa casi todos los tipos de pruebas que se definieron en la sección previa, aunque alguno predomina en cada uno de ellos.

Resumen de puntos clave 1-4

Principales contextos en que se usan las pruebas
1. Clínico
2. Educativo
3. Laboral
4. Investigación

La **primera** categoría incluye los campos de la psicología clínica, consejería, psicología escolar y neuropsicología. Consideramos todas estas aplicaciones bajo el rubro de **uso clínico**; en éstas, el psicólogo trata de ayudar a un individuo que tiene (o puede tener) algún tipo de problema, que puede ser grave (p. ej., esquizofrenia) o moderado (p. ej., elegir una carrera). Las pruebas ayudan a identificar la naturaleza y gravedad del problema, y brindan algunas sugerencias para enfrentarlo; también pueden ayudar a medir el progreso de los resultados de dichas sugerencias.

Un sinfín de investigaciones han documentado el alcance de los usos clínicos de las pruebas. Aquí ofrecemos ejemplos importantes de una selección de dichas investigaciones; para ver resúmenes de muchas de ellas, se puede consultar a Hogan (2005a). En capítulos posteriores, describiremos los usos de las pruebas específicas.

- Los psicólogos que se desempeñan en escenarios de la salud mental y hospitales estatales pasan de 15 a 18% de su tiempo en actividades de evaluación.
- Más de 80% de los neuropsicólogos pasa cinco horas semanales o más haciendo evaluaciones, y una tercera parte de ellos pasa más de 20 horas semanales en esta actividad.
- Los psicólogos escolares pasan cerca de la mitad de su tiempo laboral en actividades de evaluación.
- Una muestra de 100 informes de neuropsicología forense llevados a cabo para evaluar casos de tipos específicos de daño personal, incorporó un promedio de 12 diferentes pruebas por informe; en un informe se incluyeron hasta 32 pruebas.
- En un estudio de consejería psicológica, dos tercios informaron usar pruebas objetivas y un poco menos de un tercio declaró usar pruebas proyectivas.

Todos estos grupos emplean pruebas de inteligencia, pruebas objetivas de personalidad y técnicas proyectivas; la mayoría también usa pruebas neuropsicológicas. Los psicólogos consejeros a menudo recurren a medidas de intereses vocacionales.

Una visión general de las investigaciones muestra que las pruebas tienen un papel destacado en la práctica profesional de la psicología. Debemos agregar que en todos estos campos se requiere una capacitación especializada para aplicar e interpretar las pruebas. El trabajo a nivel doctoral en áreas como la clínica, consejería y psicología escolar por lo general supone diversos cursos completos en el uso de pruebas, lo cual va más allá del nivel introductorio de este libro.

El **segundo** uso principal de las pruebas se presenta en **escenarios educativos**, aparte del uso clínico que hace el psicólogo escolar o el consejero. Aquí nos referimos primordialmente al uso de las pruebas de capacidad y rendimiento que se aplican de manera grupal. Los usuarios reales de la información que proporcionan estas pruebas son maestros, administradores educativos, padres y público general, en especial funcionarios como legisladores y comisiones escolares. El uso de pruebas estandarizadas en escenarios educativos se resuelve en dos subdivisiones. En primer lugar, hay pruebas de rendimiento que se usan para determinar el nivel del aprendizaje del alumno. Incluso limitando nuestro recuento a las pruebas estandarizadas de rendimiento (es decir, dejando fuera un vasto conjunto de pruebas hechas por maestros), podemos decir que cada año se aplican decenas de millones de ellas. Las pruebas de rendimiento también se utilizan para documentar la competencia con el fin de obtener una certificación o licencia en muchas profesiones.

En segundo lugar, en escenarios educativos las pruebas se emplean para predecir el éxito en el trabajo académico. Los principales ejemplos de esta categoría son los exámenes de admisión que se aplican en universidades. Por ejemplo, en EUA cerca de dos millones de estudiantes hacen el SAT cada año, mientras que casi un millón hace el ACT; el *Graduate Record Examination* (GRE): General se aplica aproximadamente a 300 000 estudiantes cada año y los *Law School Admission Tests* (LSAT [Pruebas de Admisión a la Escuela de Derecho]), a cerca de 100 000. Es todavía mayor el número de alumnos de escuelas primarias y secundarias a los que se les aplican pruebas de capacidad mental de aplicación grupal como parte de los programas regulares de evaluación escolar.

La **tercera** categoría principal incluye la aplicación de pruebas al **personal de trabajo** o a los **solicitantes de empleo**. Los principales usuarios en esta categoría son las empresas y la milicia. Hay dos tareas fundamentales. La primera es elegir a los individuos más calificados para ocupar un puesto: "más calificados" significa por lo regular "con mayor probabilidad de tener éxito". Por ejemplo, tal vez queremos elegir de un conjunto de aspirantes a los individuos que tienen mayor probabilidad de tener éxito como vendedores, gerentes, secretarias o vendedores telefónicos. Hay pruebas que pueden ser útiles en el proceso de selección, como las medidas de capacidad mental general, habilidades específicas relacionadas con el trabajo o rasgos de personalidad. Desde luego, también se puede usar información que no provenga de las pruebas: cartas de recomendación y registros de empleos previos, como sucede de manera regular.

La segunda tarea en el área de empleo de personal u organizacional, tiene un escenario inicial diferente. En el primer caso, teníamos un conjunto de aspirantes, de los cuales

elegimos a los mejores, pero en el segundo caso, tenemos un grupo de individuos que serán contratados y necesitamos asignarlos a distintas tareas para optimizar la eficiencia general de la organización. Éste es un objetivo común en la milicia, donde se debe desplegar una gran cantidad de individuos. Una vez reclutados, ninguno será expulsado, sino que será empleado de un modo u otro. Las pruebas pueden aportar información útil acerca de la colocación óptima de los recursos humanos en este escenario. El *Armed Services Vocational Aptitude Battery* (ASVAB [Batería de Aptitudes Vocacionales de los Servicios Armados]) fue diseñado con este objetivo. Entre los 1000 nuevos reclutas, algunos pueden ser particularmente hábiles para las actividades mecánicas, otros en las labores de oficina y otros más en tareas de comunicación exclusivamente verbales.

La **cuarta** categoría principal del uso de pruebas, y en la que mayor diversidad existe, es la **investigación**. Las pruebas se usan en cualquier área imaginable de investigación en psicología, educación y otras ciencias sociales y de la conducta. Por comodidad, es posible identificar tres subcategorías del uso en investigación. Primero, a menudo sirven como variable dependiente. Por ejemplo, en un estudio sobre los efectos de la cafeína en la memoria a corto plazo, el *Wechsler Memory Scale* [Escala Wechsler de Memoria] puede ser la definición operacional de "memoria". En un estudio de diferencias de género en el autoconcepto, el *Piers-Harris Children's Self-Concept Scale* puede constituir la definición de autoconcepto. En un estudio longitudinal de los efectos de un programa mejorado de nutrición en el desempeño escolar, el *Stanford Achievement Test*, aplicado del segundo al sexto grado, puede servir como medida del desempeño. Hay varias ventajas importantes al usar una prueba existente como definición operacional de una variable dependiente en tales estudios. La primera es que el investigador no tiene que preocuparse por elaborar una nueva medida; la segunda es que las pruebas disponibles deben tener propiedades conocidas, como normas y confiabilidad; y la tercera y más importante es que su uso ayuda a que otros investigadores repliquen el estudio.

La segunda subcategoría del uso en la investigación consiste en la **descripción de muestras**. Las características importantes de las muestras usadas en una investigación deben delinearse. La sección del método de un artículo de investigación a menudo ofrece información acerca de la edad y género de los participantes; algunas características se describen valiéndose de información obtenida por medio de pruebas; por ejemplo, medias y desviaciones estándar de pruebas de inteligencia, rendimiento o personalidad. En un estudio con estudiantes universitarios, puede ser útil saber el promedio de las puntuaciones en el SAT o el ACT, mientras que, en uno con pacientes ancianos de un hospital público, puede ser útil saber las puntuaciones del MMPI. Nótese que en estas instancias, las puntuaciones de las pruebas no se usan como variables dependientes, sino sólo para describir las muestras de una investigación.

La tercera subcategoría consiste en la **investigación sobre las pruebas mismas**. Como veremos en el siguiente capítulo, revistas enteras se dedican a este tipo de investigación; además, la elaboración de nuevas pruebas es una empresa importante de la investigación. Ya que las pruebas tienen un papel destacado en las ciencias sociales y de la conducta, la investigación continua sobre ellas es una valiosa contribución profesional.

Temas de gran importancia: supuestos y preguntas fundamentales

Un tercer modo de introducirse en el campo de las pruebas es examinar los supuestos y preguntas fundamentales de él. Cuando los psicólogos piensan de manera cuidadosa sobre las pruebas, sin importar el tipo de prueba, ¿qué aspectos le preocupan y qué suposiciones hacen? Describir estas cuestiones y suposiciones básicas nos ayuda a entender qué se trata en este campo.

Resumen de puntos clave 1-5

Cuatro supuestos cruciales
1. Las personas difieren en rasgos importantes.
2. Podemos cuantificar estos rasgos.
3. Los rasgos son razonablemente estables.
4. Las medidas de los rasgos se relacionan con la conducta real.

Supuestos básicos

Empecemos este modo de explorar el campo identificando las suposiciones que solemos hacer, que son cuatro y, en parte, se traslapan aunque son distintas. Primero, asumimos que los seres humanos tienen **rasgos o características** reconocibles; por ejemplo, rasgos de capacidad verbal, memoria, extroversión, cordialidad, capacidad de razonamiento cuantitativo, autoestima, conocimiento de la historia de Irlanda y depresión. Además, asumimos que estos rasgos o características describen aspectos **potencialmente importantes** de las personas y, en particular, que las **diferencias** entre los individuos son potencialmente importantes. Las personas somos iguales en muchos sentidos; todos necesitamos oxígeno, pues sin él moriríamos con rapidez; así que no diferimos en ese sentido. Casi todos usamos el lenguaje en cierta medida, lo cual es una característica humana distintiva. Sin embargo, también somos diferentes de uno u otro modo; algunas personas son más altas que otras, algunas están más deprimidas que otras, algunas son más inteligentes. Asumimos que tales diferencias en los rasgos que medimos son importantes y no triviales.

¡Inténtalo!

Ya hemos nombrado diversos rasgos humanos (capacidad verbal, depresión, etc.). Trata de nombrar más de ellos, algunos en el ámbito de la capacidad y otros en el de la personalidad
Rasgos de capacidad: _____
Rasgos de personalidad: _____

Segundo, asumimos que podemos **cuantificar** estos rasgos, lo cual consiste en ubicar los objetos (en este caso, personas) a lo largo de un continuo. El continuo debe pensarse como algo que va de abajo hacia arriba o de menos a más; corresponde al rasgo que estudiamos. En su nivel más primitivo, la cuantificación implica distinguir entre los objetos en el continuo; la distinción puede ser sólo entre dos categorías etiquetadas como 0 o 1. En el siguiente nivel de sofisticación, se utiliza el concepto de "más o menos" a lo largo del continuo, como se muestra en la figura 1-1. La gente se ubica a lo largo del continuo de un rasgo. Examinamos estos conceptos de cuantificación con mayor detalle en el capítulo 3; por ahora, haremos notar el supuesto de que tal cuantificación de un rasgo es una noción fundamental en nuestro trabajo. Este supuesto de la "cuantificación" es el que da origen al uso del término **medida** en el campo de las pruebas. De hecho, en muchos contextos, "medida" se usa como sinónimo de "prueba"; por ejemplo, la pregunta "¿qué medida se usa para evaluar la inteligencia infantil?" es equivalente a "¿qué prueba se usa para evaluar la inteligencia infantil?"

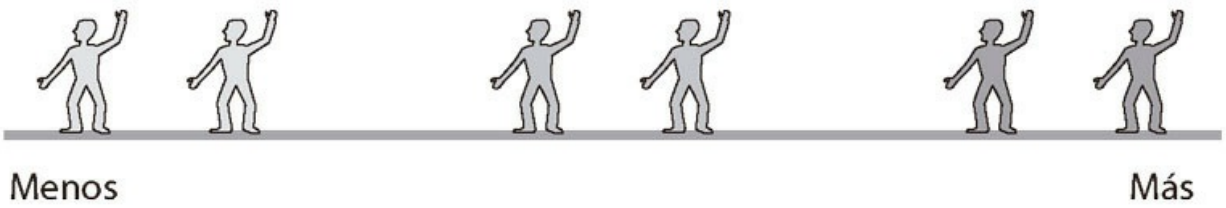


Figura 1-1. El continuo que, asumimos, existe de un rasgo.

Tercero, asumimos que los rasgos tienen cierto grado de **estabilidad o permanencia**. No es necesario que sean estables por completo, pero no deben fluctuar de manera drástica de un momento a otro. Si el rasgo mismo no es lo suficientemente estable, no importa qué tan refinada sea nuestra prueba, pues no podremos hacer mucho con ella.

Cuarto, asumimos que los rasgos estables que cuantificamos con nuestras pruebas tienen una **relación importante con la conducta** real en situaciones vitales. Desde un punto de vista teórico, este cuarto supuesto es el menos importante; es decir, como teóricos podemos estar satisfechos mostrando que podemos cuantificar un rasgo psicológico específico sin importar su relación con cualquier otra cosa. Sin embargo, desde una perspectiva práctica, este supuesto es decisivo. En términos pragmáticos, podríamos decir que no importa cuán elegante es el modo en que una prueba cuantifica un rasgo psicológico; si ésta no se relaciona con algo más, no estamos interesados en ella.

Preguntas fundamentales

Ahora consideraremos las preguntas fundamentales acerca de las pruebas. En muchos sentidos, esas preguntas se relacionan o son extensiones de los supuestos que

enumeramos antes. Anticipándonos, haremos notar que estas preguntas abordan precisamente los temas que cubren los capítulos 3, 4, 5 y 6, donde aprenderemos cómo los psicólogos tratan de responder a dichas preguntas.

Primero, preguntamos por la **confiabilidad** de la prueba, la cual se refiere a la estabilidad de sus puntuaciones. Por ejemplo, si hago la prueba hoy y mañana la hago otra vez, ¿obtendré aproximadamente la misma puntuación? Examinamos este tema con suficiente detalle en el capítulo 4. Debe notarse que esta cuestión no es exactamente la misma que tratamos en nuestro tercer supuesto, la cual se refería a la estabilidad del rasgo mismo, mientras que la confiabilidad se refiere a la estabilidad de nuestra medición del rasgo.

Segundo, preguntamos acerca de la **validez** de la prueba, la cual se refiere a lo que ésta mide en realidad. Si la prueba se propone medir la inteligencia, ¿cómo sabemos si, en efecto, mide la inteligencia? Si la prueba se propone medir la depresión, ¿cómo sabemos que mide la depresión? El área de la validez incluye el concepto de **neutralidad**, la cual es el lado frívolo del sesgo, pues se refiere a que la prueba mida de una manera equitativa en diversos grupos, por ejemplo, entre géneros, edades, grupos raciales o étnicos y distintas áreas geográficas. Esta pregunta, en el fondo, se refiere a la validez de la prueba. Nos dedicaremos con detalles a la validez en el capítulo 5 y a la neutralidad en el capítulo 6.

Tercero, preguntamos cómo interpretar las puntuaciones de una prueba. Olivia obtuvo una puntuación de 13 reactivos correctos de 20 posibles en una prueba de aritmética, ¿es una buena calificación o es mala? Pete respondió "verdadero" en 45 de 60 reactivos en una escala de depresión, ¿significa que está deprimido o eufórico? La interpretación de las puntuaciones de una prueba, por lo general, depende del uso de **normas**, las cuales se basan en las puntuaciones de grandes grupos de individuos. En el capítulo 3, describimos los tipos de normas que se usan con las pruebas y cómo se elaboran.

Las preguntas relacionadas con la confiabilidad, validez y normas son fundamentales en lo que se refiere a las pruebas, y los intentos por responderlas constituyen el meollo de la teoría de las pruebas. Éstos son los temas que nos preocupan en cualquier tipo de prueba; sin embargo, tenemos que agregar dos tipos de preguntas a nuestro catálogo de preguntas fundamentales. Saber cómo se elaboró una prueba nos ayuda a menudo a comprender mejor la confiabilidad, validez y normas; por tanto, la elaboración de las pruebas es otro tema crucial. Además, necesitamos considerar un gran número de aspectos prácticos. ¿Cuánto cuesta la prueba? ¿Cuánto tiempo se lleva? ¿Se consigue con facilidad? ¿Está disponible en otros idiomas aparte del inglés? Todas estas preguntas prácticas son importantes aunque no sean parte de la teoría de las pruebas.

Resumen de puntos clave 1-6

Preguntas fundamentales sobre las pruebas

- Confiabilidad

- Validez
- Normas
- Elaboración de la prueba
- Asuntos prácticos

Perspectiva diferencial

Como nota final, al considerar los supuestos y preguntas fundamentales, queremos llamar la atención hacia lo que denominaremos la perspectiva diferencial. En muchas áreas de las ciencias sociales y de la conducta, intentamos formular leyes o generalidades que se apliquen, más o menos, a todos; por ejemplo, ¿cuál es el programa de reforzamiento skinneriano más eficaz para aprender una habilidad?, ¿cuál es el nivel óptimo de estrés para realizar cierta tarea?, ¿el psicoanálisis cura las fobias? Formular estas preguntas sugiere que hay una respuesta que será, en general, válida para la gente. En contraste, la **perspectiva diferencial** asume que la respuesta puede diferir para distintas personas; estamos más interesados en la manera en que los individuos son diferentes que en la manera en que se parecen. Esta perspectiva diferencial permea en el campo de las pruebas; tener en mente esta perspectiva ayuda a pensar acerca de los asuntos relacionados con las pruebas.

Un debate que está surgiendo en el contexto de la perspectiva diferencial se relaciona con cómo pensamos acerca de las diferencias: como **síndromes** o como **dimensiones** (véase Widiger & Costa, 2012). Un síndrome describe un padecimiento específico, como tener un tobillo roto o un tumor cerebral: se tiene o no se tiene. El enfoque dimensional describe un continuo que va de menos a más o de abajo hacia arriba. Ejemplos muy claros son la estatura o la velocidad al correr un kilómetro, pues no son algo que se pueda o no tener. Pero ¿qué hay con la depresión? ¿Es un padecimiento específico o simplemente el extremo inferior de algún continuo? ¿Qué hay con los problemas de aprendizaje? ¿Y qué con...? Es obvio que esta lista podría continuar indefinidamente. El debate síndrome frente a dimensión tiene implicaciones importantes para nuestra manera de pensar sobre los resultados de las pruebas psicológicas.

Perspectiva histórica «12-20

Una cuarta manera de introducirse en el campo de las pruebas es examinando sus orígenes históricos. ¿Cómo se llegó al estado actual? Saber esto, a menudo, es decisivo para comprender los asuntos que se plantean en nuestros días. Primero, bosquejamos los principales períodos y eventos en la historia de las pruebas; segundo, esbozamos algunas de las fortalezas más importantes que han influido en el desarrollo del campo de las pruebas. Al armar esta historia, nos hemos valido de distintas fuentes, muchas de las cuales relatan los mismos detalles, pero desde una perspectiva un poco diferente. Para saber más de los períodos más tempranos, se puede consultar a Boring (1950), DuBois (1970), Hilgard (1987), Misiak (1961) y Murphy (1949).

La historia de este campo puede dividirse, para mayor comodidad, en **siete períodos principales** (véase el Resumen de puntos clave 1-7). La mayoría de estos períodos tiene un tema dominante que ayuda a organizar nuestra comprensión del flujo de los eventos. Trazamos límites cronológicos entre los períodos redondeando los tiempos sólo por razones pedagógicas, pero a veces traspasamos estos límites autoimpuestos para mantener la continuidad. Al bosquejar el desarrollo cronológico del campo, evitamos recitar fechas, lo cual puede adormecer la mente, porque preferimos capturar el espíritu de los diferentes períodos y transiciones entre ellos. Incluimos una juiciosa selección de fechas para destacar los eventos particularmente representativos de un periodo. El lector encontrará más fácil concentrarse en los temas que en las fechas exactas, pero es útil aprenderse de memoria algunas fechas.

Resumen de puntos clave 1-7

Principales períodos en la historia de las pruebas

1. Antecedentes remotos	Hasta 1840	
2. Creación del escenario	1840-1880	40 años
3. Raíces	1880-1915	35 años
4. Florecimiento	1915-1940	25 años
5. Consolidación	1940-1965	25 años
6. Pasado reciente	1965-2000	35 años
7. Actualidad	2000 al presente	

Antecedentes remotos: hasta 1840

El primer periodo es, en realidad, artificial; es tan extenso que casi desafía cualquier intento de hacer un resumen serio, pero necesitamos empezar en algún punto.

Identificamos tres puntos dignos de atención en esta amplia extensión de tiempo. Primero, observamos que las raíces remotas de la psicología, como de muchos otros campos, se encuentran en la filosofía. Entre los pensadores clásicos de las épocas antigua, medieval y moderna, había una clara falta de interés en el tema de las diferencias individuales o en la medición de los rasgos. Si empleamos el método moderno de "frecuencia de citación" para definir la influencia de los autores de hace 2500 años, sin duda Aristóteles, Platón y Aquino serían los tres más importantes (más allá de las sagradas escrituras). Examinar los escritos de estos tres gigantes, así como los de sus colegas, revela un interés predominante en definir qué es lo común en los seres humanos o qué es, en general, la verdad, más que las diferencias entre ellos. Consideremos, por ejemplo, el Peri Psyche (también conocido por su nombre latín De Anima, traducido al español como "Acerca del alma"). Escrito alrededor del año 350 a. de C., este trabajo se cita a menudo como el primer libro de texto de psicología; de hecho, le dio nombre al área. En el libro inicial de este tratado, Aristóteles (1935) dice: "Intentamos examinar e investigar, primero, la naturaleza y esencia del alma y, después, sus atributos [esenciales]" (p. 9). Ésta no es la sustancia de la perspectiva diferencial.

Platón, la otra gran luminaria del mundo antiguo, cuya influencia no ha disminuido, también se concentró en lo general y, aun más que Aristóteles, en lo abstracto. El escritor más influyente del periodo medieval fue Tomás de Aquino; en lo que respecta a los temas de psicología, recapituló mucho del trabajo de Aristóteles. De hecho, consideró su tarea principal reconciliar la teología cristiana con la síntesis aristotélica, por lo que Aquino adoptó el concepto de Aristóteles de capacidades humanas y manifestó el mismo desinterés en las diferencias humanas, pues prefirió concentrarse en las características generales de la naturaleza humana. Desde luego, estos filósofos no eran tontos, sino agudos observadores de la condición humana. Hicieron comentarios ocasionales –siempre fascinantes– sobre los temas de las diferencias individuales, pero siempre eran estrictamente incidentales, no el centro de atención.

Después de la época medieval, el Renacimiento atestiguó un verdadero despertar a lo individual, pero este interés se reflejó primordialmente en las producciones artísticas, la gloriosa profusión de pinturas, esculturas y construcciones que aún nos dejan sin aliento. Los pensadores dominantes del Renacimiento tardío y el periodo moderno temprano siguieron ocupándose del funcionamiento de la mente humana. Por ejemplo, Descartes, Locke, Hume y Kant plantearon preguntas –y dieron respuestas– que forman parte de los antecedentes remotos de las raíces de la psicología, pero el énfasis continuó puesto en lo que era común.

En lo que respecta al modo en que se hacían los exámenes en nuestro pasado remoto, DuBois (1970) observa que los exámenes escritos no eran comunes en la tradición educativa occidental. La práctica más usual en las escuelas en la Antigüedad, el Medioevo y, de hecho, hasta mediados del siglo XIX fue el examen oral. Los vestigios de esta práctica persisten en la actualidad en las defensas orales de las tesis de licenciatura, maestría y doctorado; en inglés, cuando alguien presenta una de estas defensas, se dice "taking your orals" [tomar tus orales], como si se tratara de algún tipo de píldora de mal

sabor (en realidad, es mucho peor que una píldora). DuBois menciona que los exámenes escritos surgieron en las notables escuelas jesuitas a finales del siglo XVI, que son las antecesoras de la red actual de escuelas secundarias, colegios y universidades jesuitas de todo el mundo. El *Ratio Studiorum* jesuita, una especie de guía curricular temprana, sentó reglas estrictas (¡estandarización!) para llevar a cabo los exámenes escritos.

Por último, algunos libros de texto informan del equivalente de los exámenes de servicio civil que se usaron de manera habitual en China desde el año 2000 a. de C. Sin embargo, Bowman (1989) sostiene de modo convincente que estos informes se basan en fuentes históricas inadecuadas (apócrifas, podríamos decir) y que la aplicación más antigua de tales pruebas ocurrió probablemente alrededor del 200 a. de C. No obstante, sea en el año 200 o 2200 a. de C., se trata de un desarrollo histórico interesante; este sistema continuó hasta principios del siglo XX y puede haber tenido cierta influencia en las pruebas del servicio civil en los países occidentales.

Creación del escenario: 1840-1880

Los hechos ocurridos entre los años 1840 y 1880 pusieron el escenario para las estrellas que habrían de ser los principales personajes del drama que se desarrolló después. Esta puesta en escena está constituida por un gran conjunto de eventos inconexos; sin embargo, en retrospectiva, podemos ver cuatro hilos que se entretejen.

Primero, a lo largo de este periodo, tanto el interés científico como la conciencia pública de la enfermedad mental aumentaron enormemente. Desde los primeros impulsos de Philippe Pinel en Francia, Samuel Tuke en Inglaterra y Benjamin Rush en EUA, surgió un gran número de intentos para mejorar el diagnóstico y tratamiento de los enfermos mentales. Dorothea Dix (figura 1-2) personifica el lado humanitario de tales intentos; alrededor de 1840 comenzó una cruzada prácticamente mundial que dio por resultado mejoras en las condiciones de cárceles y hospitales. Del lado científico, empezaron a surgir métodos para diagnosticar enfermedades mentales, incluyendo el retraso mental; por ejemplo, aparecieron métodos simples para evaluar la capacidad mental, como el tablero de formas de Seguin (Figura 1-3). Estas primeras medidas no tenían normas ni datos de confiabilidad, pero, al menos, presagiaron los elementos de las medidas que se desarrollarían más tarde.

Figura 1-2. Dorothea Dix, paladín del mejoramiento de las condiciones hospitalarias.

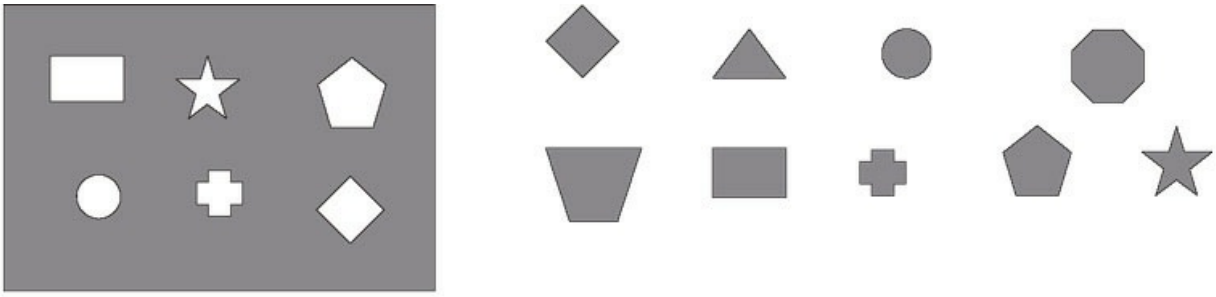


Figura 1-3. Tablero de formas como el de Seguin.
Fuente: Archivos de la Historia de la Psicología de EUA.

Un segundo desarrollo significativo de este periodo fue la adopción de exámenes formales escritos por parte del Comité Escolar de Boston –en esencia, el consejo escolar de la ciudad– bajo la dirección de Horace Mann, probablemente el educador más influyente de aquellos días. Mann defendió, no sólo en Boston sino en todo EUA, un mejoramiento sustancial en el modo en que las escuelas evaluaban a sus alumnos.

Tercero, llegó la era de Darwin. Su sorprendente obra El origen de las especies por medio de la selección natural apareció en 1859, pero quizá más importante para el incipiente campo de la psicología fueron sus siguientes libros: El origen del hombre y la selección en relación al sexo de 1871 y La expresión de las emociones en los animales y en el hombre de 1872. Desde luego, estos libros pusieron al mundo de cabeza, pero ¿por qué fueron tan importantes para la psicología? Porque hicieron que la gente pensara sobre las diferencias: primero, entre las especies y, luego, entre los individuos. En particular, hicieron que Francis Galton pensara en las diferencias individuales; hablaremos más de Galton en un momento.

Cuarto, surgió la psicología experimental. El año que tradicionalmente se considera su fecha de nacimiento es 1879, cuando Wilhelm Wundt abrió su laboratorio en la Universidad de Leipzig, ciudad alemana que actualmente cuenta con medio millón de habitantes y se ubica a unos 120 kilómetros al sur de Berlín. La psicología experimental fue una ramificación de la fisiología, cuya conexión fue la psicofísica. En sus inicios, la psicología experimental fue, en esencia, sinónimo de la psicofísica. Su contribución al campo de las pruebas, para bien o para mal, fue doble; por un lado, como toda ciencia de laboratorio, se concentró en la estandarización de las condiciones y la precisión de las mediciones. Por otro, se concentró en procesos elementales, por ejemplo, sensación, umbrales, percepción y reacciones motrices simples. El laboratorio de Wundt fue elegido por muchos psicólogos de esa época para formarse, por lo que sus intereses y métodos ejercieron gran influencia en el naciente campo de la psicología.

Así llegamos a 1880: la psicología experimental es una nueva ciencia, la evolución se discutía en todo el mundo, hay un amplio interés en la enfermedad mental, incluyendo el retraso mental, y algunos pioneros tratan de llevar la educación al terreno de la ciencia.

Raíces: 1880-1915

Las raíces del campo de las pruebas, tal como se encuentra en la actualidad, datan del periodo comprendido entre 1880 y 1915, pues las primeras medidas que tuvieron influencia duradera aparecieron en él. Muchos de los temas y métodos básicos emergieron en una forma más o menos explícita. Al principio de este periodo, había pocos –muy pocos– ejemplos que uno podría señalar y decir: ésa es una prueba. En cambio, al final de este periodo, había un ejército de instrumentos, algunos de los cuales son prácticamente idénticos a los actuales, excepto por algunas palabras arcaicas. Al principio de este periodo, el coeficiente de correlación y el concepto de confiabilidad no se habían inventado y, al final, estas piedras angulares metodológicas de las pruebas no

sólo se habían inventado, sino que se habían elaborado e incorporado a la reciente teoría de las pruebas mentales.

Destacaremos los eventos y personalidades clave de este emocionante periodo; nos centraremos en cuatro individuos y, además, mencionaremos otro personaje y luego un amplio conjunto de otros contribuidores.

La primera figura clave fue **Francis Galton** (figura 1-4), a quien muchos consideran el fundador del campo de las pruebas psicológicas. Señor británico acaudalado e independiente, nunca tuvo un trabajo real, ni siquiera como profesor universitario. Fue un diletante, pero lo fue a lo grande, con una impresionante creatividad y versatilidad.

Figura 1-4. Francis Galton: diletante extraordinario y nexo entre la teoría de la evolución y la psicología.

Digno sobrino segundo de Charles Darwin, Galton fue el primero en llevar la teoría de la evolución al naciente campo de la psicología. Su interés radicaba en la herencia, sobre todo en la herencia de altos niveles de capacidad. Él la llamó "genio" y la estudió en una amplia variedad de áreas, como la música, la milicia, el liderazgo político y la literatura.

Al tratar de examinar las relaciones entre las muchas variables que estudió, Galton inventó un gráfico de distribución bivariada; como consecuencia de esto, indujo a Karl Pearson, un matemático británico contemporáneo, a inventar el coeficiente de correlación. Galton tuvo el tiempo, los recursos y la personalidad para lograr muchas cosas, pues, además, fue un proselitista. Difundió sus ideas sobre los métodos de la medición mental; a pesar de que no ostentaba una posición de prestigio, al parecer para 1910 todos conocían su trabajo.

El principal contribuidor estadounidense para el desarrollo de las pruebas fue **James McKeen Cattell** «14. Después de un breve periodo en la Universidad de Pennsylvania, pasó la mayor parte de su carrera profesional en la Universidad Columbia en la ciudad de Nueva York. La preparación de Cattell era ideal para combinar dos corrientes metodológicas. Por un lado, hizo su trabajo de graduación primero con Wundt en Leipzig, donde refinó sus habilidades en los estudios rigurosos de laboratorio de la tradición psicofísica. Por otro, después estudió con Galton, de quien al parecer absorbió la fascinación por recolectar datos sobre las diferencias individuales en los rasgos humanos. De acuerdo con la noción predominante en ese momento, Cattell creía que la clave del funcionamiento mental eran los procesos elementales, por lo que creó una batería de 50 pruebas, de las cuales 10 eran consideradas los pilares (véase cuadro 1-1), pues cubrían áreas como agudeza sensorial, tiempo de reacción, bisección visual de una línea y juicios en intervalos cortos de tiempo. Cattell las aplicaba a grupos de estudiantes universitarios con el propósito de predecir el éxito académico –el abuelo conceptual de los actuales SAT y ACT– y persuadió a otros psicólogos para emprender proyectos similares. Las pruebas de Cattell fueron un colosal fracaso como predictores; sin embargo, su trabajo tuvo gran influencia. En un famoso artículo de 1890, acuñó el término **prueba mental** (Cattell, 1890), el cual se usó para caracterizar el campo los siguientes 50 años. Como era debido, aparecía un comentario de Galton después del artículo.

Cuadro 1-1. Lista abreviada de las 10 pruebas clave de Cattell
1. Presión del dinamómetro [fuerza del agarre]
2. Tasa del movimiento
3. Áreas de sensación
4. Presión causante de dolor
5. Menor diferencia perceptible en el peso
6. Tiempo de reacción ante los sonidos

7. Tiempo para nombrar los colores
8. Bisección de una línea de 50 cm
9. Juicios de 10 segundos de tiempo
10. Número de letras recordadas después de escucharlas una vez

¡Inténtalo!

De las pruebas que aparecen en el cuadro 1-1, ¿cuál crees que podría ser el mejor predictor del éxito académico?

Resumen de puntos clave 1-9

Personas importantes para establecer las raíces

- Francis Galton
- James McKeen Cattell
- Alfred Binet
- Charles Spearman
- Creadores de las "pruebas nuevas"

La tercera figura de este periodo que tuvo influencia fue el francés **Alfred Binet** (que se pronuncia Bă-nay']. En su sección sobre las pruebas mentales, Boring (1950) resume el tema de manera sucinta: "La década de 1880 fue el decenio de Galton en este campo; la de 1890, de Cattell, y la de 1900, de Binet"(p. 573). Binet es el verdadero padre de las pruebas de inteligencia; de manera inusual, su formación original era en leyes, pero después terminó la formación avanzada en medicina y en ciencias naturales. Durante la mayor parte de su carrera, Binet se concentró en investigar las funciones mentales; en contraste con Galton, Binet buscaba actividades mentales más holísticas, como usar palabras, encontrar relaciones o captar significados. En esa época, las escuelas parisinas querían identificar a los estudiantes que más probabilidades tenían de beneficiarse de enseñanza en escuelas especiales que de programas en escuelas regulares. Un comité, del que Binet y Theodore Simon formaban parte, se creó para elaborar un método que permitiera identificar a estos estudiantes; el resultado fue la Escala Binet-Simon,[3] que se publicó por primera vez en 1905. En 1908 y en 1911, aparecieron formas revisadas, en las que se usó el concepto de "edad mental". En el capítulo 8, examinaremos el nieto de la escala de Binet, la moderna Escala de Inteligencia Stanford-Binet.

Cuarto, también está el trabajo de Charles Spearman, otro inglés, cuyas contribuciones fueron de un carácter distinto del de los otros tres. Spearman no quería inventar nuevos tipos de pruebas o de reactivos, no emprendió ningún proyecto novedoso de recolección de datos, sino que fue un gran teórico y amo de los números. En 1904, publicó un

artículo en el que se anunció la teoría "bifactorial" de la inteligencia. De una manera extraña, apareció en el *American Journal of Psychology* y no en una publicación británica. Spearman reforzó su teoría con el método de las diferencias tetrad, la forma más antigua de la técnica estadística que con el tiempo se conoció como análisis factorial. Lo importante aquí es que se trata del primer intento de brindar una teoría empírica de la inteligencia humana y fue fuente de inspiración de nuevos métodos de la medición mental. Las teorías previas eran, en esencia, filosóficas, pero ésta era un nuevo tipo de teoría, porque se basaba en resultados de pruebas (la mayoría de tipo Galton) y se acompañaba de una nueva metodología matemática. De hecho, ésta era la sustancia de una nueva ciencia.

El elemento final para establecer las raíces de las pruebas no está identificado con tanta claridad, pues no se trató de una sola persona, sino de un grupo de personas que perseguían la misma meta y del mismo modo. Éste era el grupo de personas que construían con fervor las primeras versiones de las pruebas de rendimiento educativo. Respondían al mismo impulso que Horace Mann: la necesidad de llevar la educación al nivel del mundo de la ciencia. Si la educación había de realizarse de manera científica, entonces se requerían medidas precisas, confiables. Éste era un interés diferente del de Cattell, que era la predicción. Estos investigadores querían medidas de los resultados de la educación.

Con un entusiasmo casi evangélico, un grupo de autores de pruebas creó lo que llamaron "nuevo tipo" de pruebas de rendimiento. Su principal preocupación era la falta de confiabilidad del ensayo y los exámenes orales. Las pruebas de nuevo tipo eran tan objetivas como era posible, lo cual significaba en la práctica que los reactivos eran de opción múltiple, verdadero-falso y de llenar espacios. Estos reactivos se podían calificar de manera objetiva y eran más confiables que los de las pruebas del "viejo tipo". La literatura actual está repleta de referencias a las deficiencias de las pruebas del viejo tipo. En los debates actuales, el problema es la confiabilidad entre distintos calificadores. Muchas personas suponen que los reactivos de opción múltiple y otros parecidos se inventaron para el procesamiento masivo de pruebas mediante computadoras; nada podría ser más absurdo. Las computadoras ni siquiera existían cuando surgieron las pruebas del nuevo tipo; tampoco había escáner para leer las hojas de respuestas. Los autores que trabajaron en el campo de las pruebas de rendimiento en ese tiempo no estaban preocupados por la eficiencia en la calificación, sino que la confiabilidad de la calificación era lo que los apasionaba.

Florecimiento: 1915-1940 «16a

Desde sus humildes y más bien inconexas raíces de 1880 a 1915, el campo de las pruebas entró en un periodo de crecimiento espectacular. Al principio de este periodo, había pruebas, pero pocas estaban estandarizadas del modo en que pensamos de un instrumento actual; al final del periodo, en sólo 35 años, miles de ellas estaban disponibles. Las primeras ediciones de la gran mayoría de pruebas que hoy se usan

ampliamente nacieron durante este periodo. Esto sucedió en cada esfera del campo de las pruebas: capacidad mental, rendimiento, personalidad, intereses. El grado de actividad era vertiginoso. Ahora examinaremos algunos de estos desarrollos.

La clara demarcación entre el periodo de las raíces y el del florecimiento llega cuando las escalas de Binet atravesaron el océano Atlántico desde Francia hasta EUA. Podíamos ubicar una fecha entre 1910 y 1916, pero, sin importar cuál sea la fecha exacta elegida, lo importante es el trayecto transatlántico. El trabajo de Binet recibió atención casi de inmediato en EUA; algunas de las nuevas versiones estadounidenses fueron principalmente traducciones, de las cuales, quizá, la primera fue la de Goddard en 1910 (DuBois, 1970; Murphy, 1949). También hubo otras traducciones y adaptaciones; sin embargo, el evento definitivo fue la publicación de la Revisión Stanford de la Escala Binet en 1916, que se conoce popularmente como Stanford-Binet. Organizada por Lewis Terman de la Universidad Stanford, la Revisión Stanford incluyó nuevos reactivos (que casi duplicaron el número original), una nueva investigación de prueba, un ambicioso programa nacional para obtener normas y el uso de la razón CI: en conjunto, todo esto fue un exitazo. Un investigador que lo examinara hoy se burlaría, pero en su tiempo, fue un avión de propulsión a chorro, el primer hombre en la luna, el primer teléfono inteligente. En un lapso relativamente corto, la Stanford-Binet se convirtió en la definición por excelencia de la inteligencia humana, un pilar de la práctica clínica y, quizá, el símbolo más distintivo de la contribución de la psicología al mundo moderno. Así empezó el periodo de florecimiento.

Uno de los eventos que más influyeron en la historia de las pruebas fue el desarrollo de la primera prueba de inteligencia de aplicación grupal ampliamente usada. Esto sucedió en el contexto de los intentos de los psicólogos para ayudar a evaluar la gran cantidad de reclutas para el ejército cuando EUA entró a la Primera Guerra Mundial en 1917. Arthur Otis, como parte de su trabajo doctoral bajo la tutela de Lewis Terman (famoso por la Stanford-Binet), emprendió la creación de una forma de la Stanford-Binet para aplicarse de manera grupal. El trabajo de Otis resultó en las pruebas Army Alpha y Army Beta, versiones verbal y no verbal, respectivamente, que se aplicaron a cerca de 2 millones de miembros del ejército. En 1918, estas pruebas estuvieron disponibles para uso general como *Otis Group Intelligence Scale* [Escala de Inteligencia para Grupos de Otis]. Examinaremos un descendiente directo de esta prueba, el *Otis-Lennon School Ability Test*, en el capítulo 9.

Las pruebas Stanford-Binet y Otis establecieron el uso de una puntuación simple para representar la inteligencia. El mayor reto para esta práctica surgió en el trabajo de L. L. Thurstone (1938), quien sostenía que había (más o menos) siete dimensiones diferentes de la inteligencia humana. El trabajo de Thurstone produjo una gran cantidad de pruebas de inteligencia con puntuaciones múltiples en este periodo.

Una oleada de publicaciones en el periodo relativamente breve de 10 años, 1921-1930, estableció la preferencia por las pruebas de rendimiento del "nuevo tipo" (McCall, 1922; Odell, 1928; Ruch, 1924, 1929; Ruch & Rice, 1930; Ruch & Stoddard, 1927; Toops, 1921; Wood, 1923). Aunque la etapa previa presenció el inicio del desarrollo de una gran

cantidad de pruebas de rendimiento del "nuevo tipo", ninguna llegó a usarse de manera generalizada. La primera de estas pruebas que en verdad fue estandarizada a nivel nacional fue el Stanford Achievement Test, el cual apareció en 1923. Es interesante que uno de sus coautores fuera Lewis Terman, el principal arquitecto de la revisión en Stanford de la escala de Binet. La década de 1930 también presenció el origen de diversas baterías de rendimiento bien conocidas (p. ej., las series Metropolitan, Iowa y California), así como una multitud de pruebas de temas simples en cualquier área imaginable.

Las pruebas de personalidad, tanto objetivas como proyectivas, también florecieron en este periodo. El prototipo de los inventarios objetivos de personalidad actuales, el *Woodworth Personal Data Sheet* [Hoja de Datos Personales de Woodworth], se elaboró para ayudar en los procesos de reclutamiento del ejército en la Primera Guerra Mundial. En esencia, era una entrevista de lápiz y papel, 116 reactivos en total, que se respondían con "Sí" o "No", para detectar individuos que requerían un examen psicológico más completo. Muchos instrumentos similares surgieron después de la Primera Guerra Mundial. Los siguientes son ejemplos de la profusión de nuevas publicaciones en este periodo.

- Las manchas de tinta de Rorschach aparecieron en 1921. Para 1940, había diferentes sistemas de calificación de esta prueba.
- Strong y Kuder lanzaron su trabajo pionero sobre los inventarios de intereses vocacionales (Donnay, 1997; Zytowski, 1992). Describiremos las versiones actuales de estas pruebas en el capítulo 15.
- El MMPI (véase capítulo 13) se ideó en este tiempo, aunque no apareció sino después de finalizado este periodo.
- Thurstone y Likert intentaron por primera vez medir de manera sistemática las actitudes. Describiremos sus métodos, que aún se utilizan en nuestros días, en el capítulo 15.

Anteriormente, relatamos el evento por excelencia que dio inicio al periodo de florecimiento: la publicación de la Stanford-Binet Intelligence Scale [Escala de Inteligencia Stanford-Binet] en 1916. Un digno final para este periodo fue, quizá, la primera revisión de esta prueba en 1937; además, casi coincidió con la aparición de la *Wechsler Bellevue Intelligence Scale* [Escala de Inteligencia Wechsler-Bellevue] en 1939. David Wechsler, psicólogo clínico que trabajaba en el Hospital Bellevue de la ciudad de Nueva York, no estaba satisfecho usando la Stanford-Binet –prueba diseñada para niños– con sus pacientes adultos. Diseñó su prueba para que fuera más adecuada para los adultos.

Las primeras ediciones de tres publicaciones sirvieron como triples signos de exclamación cuando estaba por concluir este notable y fecundo periodo en la historia de las pruebas. La primera fue la publicación de la revista sumamente teórica Psychometrika en 1936; después, se publicó la revista con una orientación más pragmática *Educational*

and Psychological Measurement en 1941, y, por último, apareció la primera edición de Mental Measurements Yearbook de Oscar K. Buros en 1938. En el capítulo 2, describimos en detalle la actual edición de esta última publicación.

Consolidación: 1940-1965

Después del estallido de la actividad desde distintos frentes de 1915 a 1940, el campo de las pruebas entró en un periodo que bien puede denominarse como de consolidación o madurez, y tuvo una duración de 25 años: de 1940 a 1965. La actividad no se redujo; de hecho siguió floreciendo. Aparecieron nuevas ediciones revisadas de muchas pruebas que se crearon en el periodo anterior, pero también se elaboraron pruebas nuevas. El uso de las pruebas se extendió en la práctica clínica, las escuelas, las empresas y el ejército, así que las pruebas ya no eran el niño nuevo en la cuadra. Se aceptaron en la práctica profesional y se asumía que tendrían un papel destacado en diversos contextos. Varios eventos marcaron esta recién adquirida madurez.

Al principio de este periodo, por supuesto, la Segunda Guerra Mundial (1939-1945) era el centro de atención de todo el mundo. Las pruebas, más que ser creaciones nuevas como en la Primera Guerra Mundial, se usaron amplia y habitualmente para evaluar al personal del ejército. Psicólogos destacados, ahora entrenados en los métodos del campo de las pruebas desarrollados en el periodo anterior, llevaron a cabo estas aplicaciones. Además, los psicólogos clínicos realizaron su trabajo en el tratamiento de los daños psicológicos relacionados con la guerra empleando, en parte, las pruebas ahora disponibles.

La aparición de libros u otros documentos escritos definen, a menudo, este periodo histórico. Así, la Declaración de Independencia significó el surgimiento de una nación, aunque muchos otros eventos podrían haber sido tomados como desarrollos más importantes. Quizá, la mejor evidencia de la consolidación del campo de las pruebas en este periodo fue la aparición de numerosos libros donde se resumía el estatus del campo. Estos libros se convirtieron en clásicos justo porque pudieron brindar un resumen del pensamiento maduro acerca de las principales cuestiones del campo. Veinte años antes, digamos en 1930, no habría sido posible escribir estos libros porque el pensamiento sobre dichas cuestiones no había madurado.

Incrustado en este periodo de consolidación, hubo un lapso de seis años, 1949-1954, en el que apareció media docena de libros que pronto se convertirían en clásicos. Entre estas obras estuvieron las primeras versiones (de 1954 y 1955) de lo que serían los *Standards for Educational and Psychological Tests* [Estándares para las pruebas educativas y psicológias], una especie de biblia sobre lo mejor del conocimiento de las cuestiones técnicas relacionadas con las pruebas. Citaremos extractos de este libro a lo largo de capítulos posteriores.

En 1950, apareció el libro *Theory of Mental Tests* [Teoría de las pruebas mentales] de Harold Gulliksen, la entonces obra definitiva de la teoría psicométrica. Casi al mismo tiempo se publicaron las primeras ediciones de dos libros de texto seminales sobre el

campo de las pruebas: *Essentials of Psychological Testing* [Fundamentos de las pruebas psicológicas] de Lee Cronbach en 1949 y *Psychological Testing* [Pruebas psicológicas] de Anne Anastasi en 1954 (figura 1-5). Ambos libros aparecieron después en numerosas ediciones revisadas, pero las primeras ayudaron a definir un campo maduro de estudio. Así, el campo de las pruebas llegó a la década de 1960 con un amplio abanico de instrumentos, patrones de uso establecidos, una base teórica bien definida y publicaciones de referencia en que se resumía todo esto.

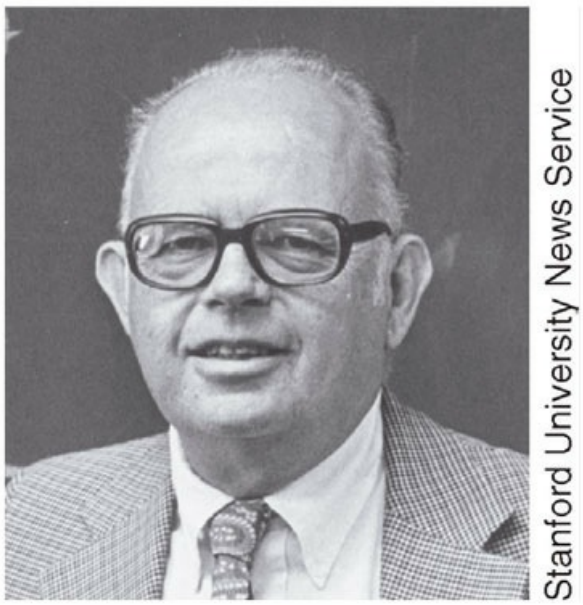


Figura 1-5. Autores de los primeros libros de texto sobre pruebas: Lee Cronbach y Anne Anastasi.

Resumen de puntos clave 1-10

Características importantes: pasado reciente

- Aparición de la teoría de la respuesta al reactivo
- Activismo legislativo y judicial
- Críticas públicas al campo de las pruebas
- Influencia de las computadoras

Pasado reciente: 1965-2000

Quien lea este libro dentro de 50 años –si es que alguien lo hace– se reirá, sin duda, del título "pasado reciente"; sin embargo, en este texto, el periodo de 35 años comprendidos entre 1965 y 2000 es "el pasado reciente". Gran parte de lo que sucedió en este periodo es simultáneo a la vida del lector de hoy, por lo que todo parece haber ocurrido apenas.

De hecho, todo esto aún se está desplegando ante nuestros propios ojos, pues se mezcla de manera imperceptible con el presente. Este periodo parece distinguirse por cuatro temas importantes.

Primero, la teoría de las pruebas ha cambiado sustancialmente. El periodo de consolidación, en esencia, resumió lo que ahora llamamos **teoría clásica de las pruebas**. A mitad de la década de 1960, apareció la **teoría de la respuesta al reactivo** o "teoría moderna de las pruebas", que consiste en un nuevo conjunto de métodos para examinar un compendio entero de temas relacionados con la confiabilidad, creación de escalas y elaboración de pruebas. El inicio del nuevo enfoque teórico quedó marcado, quizá, por la publicación de *Statistical Theories of Mental Test Scores* [Teorías estadísticas de las puntuaciones de las pruebas mentales] de Frederic Lord y Melvin Novick (1968), es decir, justo al inicio de este periodo. Este libro, anunciado como el sucesor de *Theory of Mental Tests* [Teoría de las pruebas mentales] de Gulliksen, nos introdujo en una nueva era de la teoría de las pruebas. Desde la década de 1970 hasta el presente, revistas y libros de texto se dedicaron a este campo impulsado por las aplicaciones de la teoría de la respuesta al reactivo. Nos referiremos a estos desarrollos con mayor detalle en posteriores capítulos.

Segundo, de mediados de la década de 1960 data el activismo legislativo y judicial relacionado con las pruebas y surgido principalmente, pero no exclusivamente, del gobierno federal de EUA. Hasta ese momento, el campo de las pruebas no estaba legislado, sea para bien o para mal, pero a partir de entonces la ley estableció requerimientos para algunos tipos de pruebas, mientras que otros tipos, o ciertos usos de ellos, fueron prohibidos. Los usos de las pruebas fueron cuestionados en las cortes desde muchos frentes: para clasificar estudiantes, dar empleo, graduarse, etc. Por decir lo menos, este periodo de activismo legislativo y judicial presentó un conjunto de desafíos muy particulares. Examinaremos casos específicos en este entorno en el capítulo 16.

Tercero, el campo de las pruebas se convirtió en tema de críticas públicas generalizadas en este periodo, dirigidas primordialmente a las pruebas estandarizadas de capacidad y rendimiento. Las pruebas de intereses, actitudes y personalidad salieron, en su mayoría, ilesas. En los 50 años previos a este periodo, este campo había sido visto como una herramienta científica nueva y valiosa. Para asegurarse de ello, hubo debates, pero en su mayoría habían sido como pleitos dentro de la familia, pues estuvieron confinados a las revistas especializadas en psicología y educación.

Sin embargo, a principios de la década de 1960, las críticas provenían del exterior del campo, pero se intensificaron en su interior, lo que llevó la situación mucho más allá de un pleito familiar. La lluvia de críticas tuvo tres formas principales. Primero, aparecieron varias obras populares con títulos pegajosos, de las cuales *The Brain Watchers* [Los guardianes del cerebro] (Groos, 1962) y *The Tyranny of Testing* [La tiranía de las pruebas] (Hoffman, 1962) son muy representativas. Algunas de estas críticas caían en lo puramente quisquilloso; se derramaron litros de tinta sobre las 46 diferentes maneras en que se podía interpretar un determinado reactivo de opción múltiple. Esto estaba dirigido al público general, pero era una molestia constante para los autores y editores de las

pruebas, aunque no tuviera un efecto práctico en la creación o uso de ellas. Sin embargo, esto dio origen a preguntas en la conciencia pública sobre la validez de las pruebas. El segundo tipo de crítica estaba relacionado con las diferencias étnicas y raciales en las puntuaciones de las pruebas, lo que, en realidad, era parte de los temas legales y judiciales antes mencionados. Este tipo de críticas propició una gran cantidad de investigación sobre los sesgos de las pruebas, de lo cual hablaremos en detalle en los capítulos 6 y 16. El tercer tipo de crítica fue demoledor; en esencia, decía que las pruebas indagaban en las cosas equivocadas, pues las pruebas de rendimiento de opción múltiple pasaban por alto dimensiones importantes del desarrollo del alumno. Peor aún, se decía que el uso de tales pruebas promovía los hábitos de enseñanza y aprendizaje equivocados. Para quien guste de la polémica, ¡el campo de las pruebas es el lugar ideal en la actualidad! En los capítulos posteriores, podremos revisar todas estas críticas y veremos cómo analizarlas.

Cuarto, la influencia de las computadoras ha permeado en el campo contemporáneo de las pruebas. Para el lector actual, puede ser una sorpresa que tal influencia sea reciente. Las raíces, el florecimiento y la consolidación del campo precedieron la era de la computadora; sin embargo, en los últimos 30 o 40 años la práctica de este campo ha cambiado mucho a causa de las computadoras, pero reservaremos esta historia para la discusión de las principales fuerzas que han influido en el campo de las pruebas en la siguiente sección.

Actualidad: de 2000 al presente

Decir que estamos escribiendo una historia del presente es un oxímoron, por cierto, muy peligroso. Considerar de manera errónea una irregularidad temporal como tendencia importante, o tener como punto ciego el surgimiento de una tendencia en verdad significativa, puede hacernos pasar por tontos. No obstante, concluiremos este bosquejo de la historia del campo de las pruebas identificando lo que parecen ser desarrollos destacados en la escena actual. Identificamos cinco desarrollos; a manera de prólogo de la discusión, haremos notar que las cuatro tendencias del periodo previo todavía están más que vivas y que lo que clasificamos como desarrollos actuales son frutos de él.

Primero, hay un aumento explosivo del número y diversidad de pruebas; cada día se anuncian nuevos instrumentos o revisiones de las que ya existen. Este fenómeno de crecimiento parece afectar todas las áreas de las pruebas psicológicas y educativas; una notable subdivisión de la actividad se presenta en los programas estatales de evaluación, llevados a cabo primordialmente por la reciente *No Child Left Behind Act* [Ley para que ningún niño se quede atrás] (NCLB), implementada en 2002. Como resultado, ahora cada estado es, en efecto, creador y editor de pruebas, pero también estamos siendo testigos de una profusión de nuevas pruebas de personalidad, varios trastornos y capacidades mentales. Incluso catalogar todas las entradas nuevas se ha convertido en una tarea casi imposible, y evaluarlas de manera oportuna constituye un desafío desalentador.

Este crecimiento vertiginoso resalta la necesidad de la eficiencia en el uso de las fuentes de información sobre las pruebas, que es precisamente el tema de nuestro siguiente capítulo. También resalta la necesidad de ser competente para evaluar la plétora de pruebas nuevas, a lo cual aspiramos en los capítulos 3 al 6.

Segundo, la influencia del énfasis en el manejo cuidadoso o manejo responsable se ha generalizado; desde luego, ésta no empezó en 2000, pero pasó desapercibida hasta los últimos años del periodo anterior. Ahora es una de las fortalezas con mayor influencia en la práctica clínica. El manejo responsable ejerce presión sobre el campo de las pruebas de distintas maneras. Exige pruebas más focalizadas: no uses una batería general de dos horas si una prueba más focalizada de 15 minutos es suficiente. El manejo responsable también demanda vínculos más cuidadosos entre el diagnóstico y el tratamiento y, por otro lado, entre el tratamiento y los resultados. Por lo tanto, los resultados de las pruebas deben indicar el tratamiento, y los resultados de éste deben estar documentados. En la práctica, esto implica el uso repetido de una prueba para mostrar las mejorías, definidas como un cambio en la puntuación de la prueba.

Tercero, una extensión del modelo científico-practicante que se describe más adelante es la **práctica basada en la evidencia** (PBE): la noción de que cualquier cosa que haga el psicólogo en su práctica debe estar basada en evidencia sólida. Como señalaron Norcross et al. (2008), "desde los primeros años de la década de 1990, hemos sido testigos de un crecimiento exponencial del número de artículos que hablan de la PBE. En verdad, la PBE se ha convertido en un coloso internacional" (p. 2). Las pruebas psicológicas tienen un papel decisivo en la PBE, pues mucha de su "evidencia" proviene de ellas, como las que revisaremos más adelante en este libro. Además, comprender la evidencia requiere precisamente la clase de conocimiento que exponemos en los siguientes cinco capítulos.

Las áreas cuarta y quinta del desarrollo se relacionan con las computadoras, pero de un modo diferente. La siguiente sección (Principales fortalezas) rastrea la influencia a largo plazo de las computadoras sobre el campo de las pruebas. Ahí bosquejamos estos desarrollos tardíos; sólo señalamos aquí que ellos se relacionan con el gran aumento en la aplicación y los informes en línea de las pruebas y con el desarrollo de programas de cómputo que simulan el juicio humano en el análisis de las respuestas a las pruebas. Estos dos desarrollos recientes están revolucionando ciertos aspectos del campo.

Fortalezas principales

Hay una alternativa distinta de la cronológica para ver la historia de las pruebas. Podemos examinar las fortalezas principales, tendencias o temas recurrentes que ayudaron a crear este campo y lo trajeron hasta el presente. Es una aproximación más arriesgada, porque se puede pasar por alto una tendencia importante o juzgar mal la influencia de un hecho determinado, de modo que nos convertiríamos en presa fácil para la crítica. Es difícil pasar por alto un periodo cronológico. Sin embargo, para ofrecer a los nuevos estudiantes material para la reflexión, esta segunda aproximación puede ser más beneficiosa; por ello, tomaremos el riesgo. Aquí identificamos **seis fortalezas principales** que han moldeado el campo de las pruebas como lo conocemos hoy.

Resumen de puntos clave 1-11

Principales fortalezas en la historia de las pruebas

- Impulso científico
- Preocupación por el individuo
- Aplicaciones prácticas
- Metodología estadística
- Ascenso de la psicología clínica
- Computadoras

Impulso científico

La primera fortaleza que influyó en el desarrollo de las pruebas es el impulso científico, que ha predominado a lo largo de toda su historia. Los escritos de Galton, E. L. Thorndike, Cattell, Binet y otros fundadores están llenos de referencias a la necesidad de contar con mediciones científicas. También los educadores esperaban que el desarrollo y aplicación de las pruebas del "nuevo tipo" dieran a la tarea educativa el estatus de científica. El subtítulo del artículo de Binet y Simon de 1905 se refiere a "la necesidad de hacer diagnósticos científicos". La primera oración de la introducción de Thorndike (1904) a la teoría de la medición mental establece que "la experiencia ha mostrado de manera suficiente que los hechos de la naturaleza humana pueden convertirse en material de la ciencia cuantitativa" (p. V). Esta preocupación de ser científico, junto con la de la confiabilidad entre calificadores, motivó la elaboración de las primeras pruebas de rendimiento. Por último, el campo de la psicología clínica, que trataremos de manera más exhaustiva y que ha sido uno de los campos primarios de aplicación de las pruebas, ha proclamado con firmeza su lealtad a la aproximación científica. Muchas otras profesiones, como medicina, derecho y trabajo social, han hecho hincapié en la práctica;

sin embargo, la psicología clínica siempre ha sostenido que es parte ciencia y parte práctica, y se ha valido de lo que el campo llama el **modelo científico-practicante** (véase Jones & Mehr, 2007).

Preocupación por el individuo

El campo de las pruebas ha crecido alrededor de un fuerte interés en el individuo. Esta orientación es, quizá, inevitable, ya que en este campo se trata con las diferencias individuales, lo cual es parte de la "perspectiva diferencial" que mencionamos antes. Recordemos que uno de los hilos en los antecedentes inmediatos para establecer las raíces fue el fuerte aumento de la preocupación por el bienestar de los enfermos mentales. Muchas, aunque no todas, de las aplicaciones prácticas de las que hablaremos más adelante se relacionan con la preocupación por el individuo. El trabajo de Binet buscaba identificar individuos que se beneficiaran más de las escuelas especiales que de las regulares; la primera prueba de Wechsler intentó brindar una medida más imparcial de la inteligencia de los adultos. Los SAT originales pretendían eliminar o minimizar cualquier desventaja que los estudiantes de las escuelas menos acaudaladas pudieran tener al entrar a la universidad. Las medidas de los intereses vocacionales buscaban ayudar a individualizar la selección para el trabajo. Al leer una selección representativa de manuales de pruebas y de la literatura profesional del campo a lo largo de la historia, es impresionante encontrar frecuentes referencias al mejoramiento de muchos individuos.

Aplicaciones prácticas

Cualquier desarrollo importante en el campo de las pruebas fue resultado del trabajo sobre un problema práctico. Binet trató de resolver problemas de este tipo para las escuelas parisinas; Wechsler quería una prueba mejor para sus pacientes adultos de la clínica. El MMPI pretendía ayudar en el diagnóstico de los pacientes de un hospital. El SAT surgió como tarea en la que colaboraron varias universidades para seleccionar estudiantes con distintas experiencias escolares. Los prototipos de las primeras pruebas de inteligencia de aplicación grupal e inventarios de personalidad se desarrollaron por la necesidad de procesar una gran cantidad de personal militar en la Primera Guerra Mundial. Para estar seguros, podemos encontrar excepciones a este patrón; en algunos casos se obtuvieron resultados notables de las consideraciones teóricas. Sin embargo, el patrón general parece muy claro: las pruebas se desarrollaron en respuesta a necesidades prácticas. A quien gusta del lado aplicado de la psicología le gustará el campo de las pruebas. Aunque nuestros capítulos 3 al 6 pueden parecer abstractos, los temas que se tratan ahí surgen de las necesidades prácticas.

Metodología estadística

El desarrollo del campo de las pruebas tiene una intrigante relación interactiva con el de los métodos estadísticos. Por lo común se piensa que es una relación en un solo sentido: el campo de las pruebas toma prestados los métodos de la estadística. Sin embargo, varios métodos estadísticos se crearon en respuesta a los desarrollos en este campo y, después, se adoptaron en otras áreas. El primer ejemplo fue la exposición de resultados bivariados inventados por Galton. Para promover este trabajo, Galton indujo al matemático británico Karl Pearson a crear un coeficiente de correlación; después, Spearman inventó su versión, la correlación por rangos ordenados. Más importante, al formular su teoría de la inteligencia, Spearman ideó el método de la diferencia tetrad, el abuelo conceptual del análisis factorial moderno, el cual fue un gran adelanto que llegó junto con el trabajo de Thurstone sobre las capacidades mentales básicas. Muchas de las subsiguientes elaboraciones del análisis factorial resultaron de la guerra de palabras que siguió entre Spearman y Thurstone y de los datos sobre la naturaleza fundamental de la inteligencia, un campo de batalla que aún sigue activo. Así, la historia de las pruebas ha ido de la mano con la historia de, al menos, ciertos métodos estadísticos.

Ascenso de la psicología clínica

La psicología clínica es una de las principales áreas de aplicación de las pruebas y también de la psicología. Esto es especialmente cierto si interpretamos el término **clínico** como un concepto amplio que incluye la consejería psicológica, la psicología escolar y el lado aplicado de la neuropsicología. Por un lado, las personas en la práctica clínica han requerido, presionado y ayudado para crear una plétora de pruebas; por otro, conforme han aparecido nuevas pruebas, quienes están en la práctica clínica las han utilizado. Los interesados en la historia breve de la psicología clínica pueden consultar a Routh y Resiman (2003), Trull y Prinstein (2013, en especial el capítulo 2) y Tryon (2008).

Por lo general, se atribuye a Lightner Witmer la fundación de la psicología clínica; como muchos psicólogos de su tiempo, Witmer recibió un entrenamiento avanzado con Wundt, por lo que se empapó en la metodología de la psicofísica. Al aceptar tratar a la famosa gente con "mala ortografía crónica" –lo cual, sin duda, hoy llamaríamos problemas de aprendizaje– llevó los métodos de la psicología de laboratorio al tratamiento de casos específicos. Esto ocurrió en 1896; después, Witmer abrió la primera clínica psicológica, impartió el primer curso de psicología clínica (ambos en la Universidad de Pennsylvania) y fundó una revista dedicada a la psicología clínica. Sin importar las pruebas disponibles en su tiempo ni cuán rudimentarias fueran, guiaron el diagnóstico y el tratamiento.

La historia temprana de la psicología clínica se lee de manera muy parecida a la de la historia temprana de las pruebas: la de Binet, la de Rorschach, etc. Conforme aparecieron nuevas pruebas, los clínicos las usaron; en muchos casos, era el clínico quien elaboraba las pruebas, además de que también participaron activamente en la milicia durante la Primera y Segunda Guerras Mundiales; después de la segunda, el gobierno federal hizo una fuerte inversión en el entrenamiento de psicólogos clínicos. Esto llevó a

un crecimiento explosivo de la profesión, lo cual también provocó un crecimiento en el uso de las pruebas y la necesidad de otras más nuevas. Esta relación recíproca continúa el día de hoy; en términos absolutos de pruebas empleadas, el campo de la educación encabeza el campo de las pruebas. En términos de la vertiginosa serie de varios tipos de pruebas disponibles en la actualidad, el campo clínico (entendido ampliamente) ha sido el de mayor influencia.

Computadoras

Las computadoras han tenido una profunda influencia en el desarrollo de la pruebas; como se señaló anteriormente, éste es un fenómeno muy reciente. La computadora electrónica se inventó en 1946 y se empezó a comercializar en 1951. Sin embargo, el uso de las computadoras no se generalizó sino hasta la década de 1960. Los modelos de escritorio aparecieron alrededor de 1980 y proliferaron a mediados de esa década. Así, prácticamente todas las aplicaciones de la tecnología computacional para las pruebas ocurrieron sólo en la etapa histórica más reciente de las que bosquejamos antes.

Para relatar la historia del efecto de las computadoras en el campo de las pruebas, necesitamos primero distinguir entre el escáner y la computadora, pues hay mucha confusión en este punto. Un **escáner** es un dispositivo eléctrico o electrónico que cuenta las marcas en la hoja de respuestas de una prueba, por lo que a veces se le llama escáner o lector con sensor para las marcas. A pesar de la referencia popular a las "hojas de respuesta computarizadas", éstas no se meten en una computadora, sino en un escáner. El resultado del escáner puede (o no) ser introducido en una computadora.

El primer escáner que se utilizó fue el IBM 805, que apareció en 1937; era un objeto abultado, como del tamaño de un escritorio de oficina. Un empleado insertaba las hojas de respuestas una por una en una ranura de la máquina, que funcionaba contando los circuitos eléctricos completados. No se usaba ninguna computadora; de hecho, no se habían inventado. El IBM 805 era, por mucho, más eficiente y exacto que contar a mano un registro de la prueba. Para los estándares de hoy, desde luego, es un verdadero dinosaurio.

El escáner típico de hoy (véase el ejemplo de la figura 1-6) funciona contando haces de luz, sean transmitidos o reflejados; lanza haces de luz en los espacios objetivo en la hoja de respuestas, mientras una clave le indica al escáner dónde buscar las marcas. Cualquier marca oscura –lápiz, bolígrafo– servirá. El escáner cuenta los haces de luz que atraviesan la hoja o rebotan en las marcas oscuras.

Figura 1-6. Escáner moderno pequeño; escáner lector de marcas ópticas NCS Pearson OpScan®.
Cortesía de NCS Pearson

Los modelos pequeños de escáner ahora están disponibles en la mayoría de escuelas, universidades, clínicas, negocios y otras instituciones. Las versiones potentes que usan los editores de pruebas, las agencias de gobierno y las corporaciones especializadas en escanear pueden procesar 150 hojas de respuesta por minuto (9000/hr). El resultado básico del escáner sigue siendo el mismo: el conteo o registro de marcas. En algunos casos, el conteo se imprime simplemente en un documento de respuestas; en otros casos, junto con la ubicación de cada marca, se convierten en algún medio electrónico, que puede procesarse en una computadora.

¿Las hojas de respuesta y el escáner serán los dinosaurios del mañana? Es muy probable, pues los examinados ya responden directamente en la computadora y ahí trabajamos con ellas. Esto se hace en los salones de clase, clínicas, departamentos de recursos humanos e incluso en centros ubicados en centros comerciales. Además, con las mejoras en el reconocimiento de voz, las respuestas a las preguntas de la prueba (p. ej.,

realizadas en líneas telefónicas) ahora pueden ser orales. Las palabras habladas pueden descodificarse y calificarse comparándolas con plantillas de las respuestas aceptables. Sin duda, maravillas tecnológicas adicionales aún están por venir.

Ahora, pasemos a las computadoras. Hay tres aspectos principales en la relación entre computadoras y pruebas que tienen una secuencia histórica, pero difieren más en carácter que en el orden cronológico. Además, aunque hay una secuencia histórica, es acumulativa; es decir, una vez que comenzaba una fase, ésta se quedaba en el campo. Más que reemplazarla, una nueva fase se añadía a la anterior.

Resumen de puntos clave 1-12

Relación computadora-pruebas

- Procesamiento estadístico
- Informe de puntuaciones
- Aplicación de la prueba

En la **primera** fase, las computadoras eran simples auxiliares en el **procesamiento estadístico** de la investigación de las pruebas; esta fase empezó casi tan pronto como las computadoras estuvieron disponibles comercialmente. Esto fue de gran ayuda para el campo de las pruebas, ya que permitió que se realizaran programas de investigación a gran escala y se emplearan de manera habitual metodologías sofisticadas. Este tipo de desarrollo continúa a un ritmo vertiginoso en la actualidad. Las computadoras de escritorio permiten a casi cualquier investigador realizar análisis en unos pocos segundos que antes le habrían llevado meses a equipos enteros de investigadores. En ejercicios posteriores de este libro, podrás realizar con facilidad análisis computacionales que habrían constituido una tesis de maestría hace 40 años.

En la **segunda** fase, las computadoras preparaban los **informes de las puntuaciones de la prueba**. Esto empezó con informes muy sencillos, útiles en especial para programas a gran escala. Un programa de cómputo determinaba las puntuaciones naturales (es decir, el número de respuestas correctas) empleando la información de un escáner, convertía las puntuaciones naturales puntuaciones normativas e imprimía listas de los nombres de los examinados y sus puntuaciones. Antes, todas estas operaciones se tenían que hacer a mano. Los informes impresos de la computadora se generalizaron a principios de la década de 1960; para los estándares de hoy, tales informes son primitivos: sólo usaban letras mayúsculas, todas del mismo tamaño y tipo, copiadas con papel carbón, etc. Pero en su momento, fueron una maravilla.

Las últimas etapas de esta fase evolucionaron de manera natural desde los primeros reportes sencillos. Los creadores de pruebas adquirieron maestría en la habilidad de programación y también vieron posibilidades creativas en los informes de computadora. A principios de la década de 1970, hubo una profusión de informes hechos por la

computadora cada vez más elaborados. Las capacidades de las impresoras también aumentaron y aparecieron gráficas y variaciones en las fuentes de las letras acompañadas, ahora, de información numérica, lo básico de los primeros informes.

La preparación de **informes interpretativos** fue un desarrollo importante en esta fase. Los informes del desempeño en las pruebas ya no estaban limitados a los números, sino que ahora podían describirse con palabras simples o, incluso, una narración continua, como si hubiera sido escrito por un psicólogo profesional. En el capítulo 3, explicaremos cómo se preparan tales informes y ofreceremos ejemplos de informes reales a lo largo de los capítulos posteriores.

Los desarrollos en estas líneas continúan en la actualidad. Algunos lectores de este texto probablemente contribuyan a los nuevos y mejorados informes de este tipo.

La **tercera** fase se relaciona con la **aplicación de la prueba** por medio de la computadora. Desde la perspectiva de los creadores de pruebas, hay dos tipos diferentes de aplicación mediante la computadora. El examinado puede no darse cuenta de la diferencia. El primer tipo, aplicación de la prueba basada en la computadora, presenta simplemente en la pantalla de una computadora (el monitor) las preguntas tal y como aparecerían en un cuadernillo impreso. Sólo se ponen las preguntas de la prueba en un archivo de texto y aparecerán de pronto en la pantalla. El examinado responde con ayuda del teclado; no hay nada extraordinario aquí, pero no se necesita papel.

El segundo tipo, la **aplicación adaptable por computadora**, es revolucionario. Aquí, la computadora no sólo presenta los reactivos, sino que también elige el siguiente con base en las respuestas previas del examinado. Podemos ilustrar esto con una prueba de aritmética; la pantalla presenta un reactivo, digamos 26×179. Si la respuesta es correcta, la computadora elige un reactivo más difícil de su "banco" de reactivos, digamos $1372 \div 86$. Si la respuesta es correcta otra vez, la computadora presentará otro reactivo aún más difícil. Pero si la primera respuesta es incorrecta, la computadora elegirá un reactivo más fácil, digamos $38 + 109$. Los reactivos varían en dificultad hasta que la computadora "decide" que tiene un cálculo muy exacto del nivel de capacidad aritmética del examinado. Por eso, la aplicación adaptable por computadora también se ha llamado aplicación a la medida: la selección de reactivos se hace a la medida, como un traje, del examinado. Como puede imaginarse, se necesita mucha investigación sobre el desarrollo de la aplicación adaptable por computadora para que pueda funcionar apropiadamente. La investigación depende en gran medida de la teoría de la respuesta al reactivo antes mencionada. Ésta es una de las áreas de más rápido crecimiento en el campo de las pruebas; los interesados en la descripción de los métodos pueden consultar a Parshall et al. (2002) y van der Linden y Glas (2010). En los siguientes capítulos, exploraremos algunas de las teorías y métodos que subyacen en la aplicación adaptable por computadora.

La tercera fase de las aplicaciones de la computadora está entrando a un nuevo escenario. En el pasado reciente, la resolución en línea de las pruebas se ha hecho común; en un campo, la evaluación de los intereses vocacionales, responder los inventarios e informar las puntuaciones se están convirtiendo en lo estándar. Otras áreas

de aplicación no están muy atrás; el tema principal aquí no es tanto la resolución de la prueba, sino la entrega de la información de la prueba (informes, algunos muy elaborados) a individuos sin ninguna capacitación para interpretar esa información y posiblemente sin acceso a la asesoría profesional. Como veremos en capítulos posteriores, interpretar la información de las pruebas no es siempre un asunto sencillo; la psicología siempre ha hecho hincapié en la necesidad de una capacitación adecuada para hacerlo. Los informes en línea, ajenos a la aplicación de la prueba, crean un escenario nuevo por completo. Para tener un panorama general de los temas que están surgiendo, se puede consultar a Naglieri et al. (2004).

Por último, ha surgido una aplicación de lo que se denomina **calificación automatizada** «23a, lo que significa que se ha desarrollado un programa de cómputo para simular el juicio humano al calificar productos como ensayos, planos arquitectónicos o diagnósticos médicos. Tomemos, por ejemplo, un ensayo escrito por un alumno universitario. Por lo general, sería calificado

por un miembro del cuerpo docente, pero en casos muy importantes, dos o tres de ellos pueden hacerlo y luego se promedian las calificaciones. Con la calificación automatizada, ¡un programa de cómputo califica el ensayo! Tales programas, introducidos hace muy pocos años, se usan ahora en lugar de "calificadores humanos" para calificar ensayos y otros productos en programas a gran escala de evaluación. Aquí sólo haremos notar que éste es un "nuevo juego de pelota", que probablemente verá un gran crecimiento en la siguiente década con cada vez más áreas de aplicación. Por ejemplo, ¿qué hay de la calificación por computadora de las respuestas a la prueba de manchas de tinta de Rorschach? Para saber más de este tema, se puede consultar a Dikli (2006) y Drasgow, Luecht y Bennett (2004).

Palabras finales sobre las fortalezas

En esta sección hemos incluido sólo los hechos que influyeron en la mayoría de los tipos de pruebas, si no en todas. Pero han habido otros hechos y tendencias más restringidas a un tipo de pruebas o a pocos de ellos; por ejemplo, la psicología cognitiva ha afectado las pruebas de inteligencia. El movimiento de manejo responsable ha afectado las pruebas de rendimiento. Reservamos el tratamiento de estas influencias más restringidas a los capítulos sobre los tipos específicos de pruebas, por ejemplo, la psicología cognitiva en el capítulo 7 y el manejo responsable en el 11.

¡Inténtalo!

Para ver cómo es una prueba adaptable por computadora, visita este sitio: http://edres.org/scripts/cat Haz clic en "Let's Get Started" [Empezar] y luego sigue las instrucciones. No necesitas hacer la prueba completa, sino que basta con responder algunos reactivos para que tengas una idea.

Marca esta dirección como sitio favorito, pues la usarás otra vez más adelante.

Definición

El último tema para introducirse en el campo de las pruebas es el de la definición. ¿A qué nos referimos exactamente con el término **prueba**? Desde una perspectiva estrictamente académica, éste sería el primer tema a tratar para introducirse en el campo; sin embargo, también es un comienzo árido y aburrido. Así, preferimos llegar por otros caminos. Quizá más importante, habiendo considerado los otros tremas que hemos tratado hasta aquí, estamos en una mejor posición para reflexionar sobre algunas definiciones alternativas y apreciar las diferencias en varias fuentes.

Encontrar consenso acerca de la definición de "prueba" resulta ser sorprendentemente difícil, pues esta palabra se ha usado de muchas maneras y en distintas fuentes; incluso centrándonos en las pruebas psicológicas, donde pueden resaltar distintos aspectos. Muchas definiciones son circulares, pues dicen que una prueba es lo que se usa al aplicar una prueba: en verdad inútil. No obstante, para guiar nuestro pensamiento más adelante, intentaremos abstraer de varias fuentes lo que parecen ser los elementos clave. Parece haber seis elementos en común en lo que queremos decir con "prueba" en el contexto de las ciencias conductuales.

Primero, una prueba es una especie de **procedimiento o dispositivo**. Todos concuerdan en este punto, pero puede ser útil para nosotros agregar que brinda información. Aunque quizá es demasiado obvio formularlo así, nos ayudará en discusiones posteriores. Por tanto, agregamos esto como el segundo punto: una prueba ofrece **información**. Tercero, el procedimiento o dispositivo ofrece información sobre la **conducta**; este aspecto de la definición es lo que separa una prueba de, digamos, la medición física, como la altura o el peso, o de las pruebas médicas como las que se emplean para detectar un padecimiento viral. En tiempos anteriores, de orientación conductista, la "conducta" era entendida de manera estrecha, pues incluía sólo la conducta observable externa; pero en el medio de la orientación cognitiva actual, entendemos el término de una manera más amplia para incluir los procesos cognitivos. De hecho, para hacer esto explícito, ampliaremos el objeto de las pruebas para incluir **la conducta y los procesos cognitivos**.

Cuarto, muchas definiciones hacen hincapié en que una prueba ofrece información sólo de una **muestra** de la conducta. Al aplicar pruebas, por lo general no hacemos un censo exhaustivo de toda la conducta o los procesos cognitivos de una persona, sino sólo tomamos una pequeña muestra. Esta noción será crucial cuando consideremos la confiabilidad y la validez. Quinto, una prueba es un procedimiento **estandarizado y sistemático**. Ésta es una de las características más distintivas de una prueba, pues la distingue de fuentes de información como las entrevistas informales o las observaciones anecdóticas, las cuales pueden ser fuentes de información útiles, pero no son pruebas.

Resumen de puntos clave 1-13

Elementos de la definición de "prueba"

- Proceso o dispositivo
- Ofrece información
- Conducta o procesos cognitivos
- Muestra de...
- Estandarizado
- Cuantificado

Ahora necesitamos hacer una digresión para poner en claro una cuestión terminológica que puede ser confusa. Hay **tres** usos del término **estandarizado** en el campo de las pruebas. Primero, cuando se usa en la definición de las pruebas, se refiere a procedimientos uniformes para aplicarlas y calificarlas. Hay métodos inequívocos, especificados con claridad para aplicar la prueba, y hay reglas para calificarla, y es fundamental que la prueba se aplique y califique de acuerdo con dichos procedimientos. Segundo, en otros contextos, estandarizado significa que la prueba tiene normas; por ejemplo, las normas nacionales basadas en miles de casos. De hecho, el proceso de recolectar los datos normativos a menudo se menciona como programa de estandarización de una prueba. Es claro que éste es un significado diferente al primero; se puede tener una prueba con instrucciones y procedimientos de calificación determinados sin tener ningún tipo de normas. Un tercer significado, que se encuentra sobre todo en los medios de información y en las discusiones públicas, considera equivalentes las pruebas estandarizadas y las pruebas de capacidad y rendimiento de aplicación grupal, calificadas con máquinas, de opción múltiple. Por ejemplo, el encabezado de un periódico puede informar: "Estudiantes locales mejoran en las pruebas estandarizadas" o "Haciendo trampa en supuestas pruebas estandarizadas". O un amigo puede decir, refiriéndose al desempeño en las pruebas de admisión universitarias SAT o ACT, "No me va muy bien en las pruebas estandarizadas". Este tercer significado es obviamente mucho más limitado que cualquiera de los otros dos. Es importante que el estudiante de psicología distinga entre estos tres significados del término **estandarizado**.

Un sexto y último elemento en las distintas definiciones es cierta referencia a la **cuantificación o medición**. Es decir, al final presentamos la información en forma numérica. Este elemento es muy explícito en algunas fuentes y parece estar implicado en las otras; la cuantificación puede ocurrir de una manera muy rudimentaria o muy sofisticada. Por ejemplo, una cuantificación cruda puede implicar formar dos grupos (de deprimidos y no deprimidos o de competentes y no competentes). Una medición más sofisticada puede implicar una escala cuidadosa parecida a la que se usa para medir la estatura o el peso.

Varias fuentes difieren en un aspecto de la definición de "prueba", a saber, el grado en que la prueba es evaluativa. Algunas definiciones se detienen con la información; otras incluyen referencias a una dimensión evaluativa, una inferencia o conclusión derivada de la información. Algunos libros tratan este punto distinguiendo entre los términos **prueba**,

evaluación y **valoración**. Por ejemplo, algunos autores suponen que hay diferencias entre estos tres enunciados: aplicamos una prueba de inteligencia a Abigail, evaluamos la inteligencia de Abigail o valoramos la inteligencia de Abigail. Sin embargo, en muchas fuentes, estos tres términos son intercambiables. Por ejemplo, los estándares para pruebas psicológicas (AERA et al., 1999) parecen combinar los tres términos definiendo "prueba" como "un dispositivo evaluativo" (p. 183) y "aplicación de pruebas" como "cualquier procedimiento... para evaluar" (p. 180); estas definiciones se mantienen en la edición de 2013 de dichos estandáres. Nosotros no tomamos una posición definitiva sobre este asunto; simplemente hacemos notar que distintas fuentes lo tratan de manera diferente.

A partir de la discusión anterior, formulamos la siguiente definición: **Una prueba es un proceso o dispositivo estandarizado que ofrece información sobre una muestra de la conducta o los procesos cognitivos de una manera cuantificada.**

Resumen

1. Clasificamos las pruebas en cinco categorías principales: pruebas de capacidad mental, de rendimiento, de personalidad, de intereses y neuropsicológicas. Cada categoría se divide en otras subcategorías. Utiliza este acrónimo para recordar estas categorías: MERPIN.
2. Las pruebas también se pueden caracterizar de acuerdo con las siguientes características: a) son de lápiz y papel o de ejecución, b) de velocidad o de poder, c) de aplicación individual o grupal, d) de ejecución máxima o típica y e) su interpretación depende de una norma o de un criterio.
3. Los usos principales de las pruebas son en las áreas clínica, educativa, laboral y de investigación.
4. Cuatro supuestos importantes apuntalan la empresa de las pruebas:

- La gente tiene diferencias en sus rasgos, las cuales son importantes.
- Podemos cuantificar estos rasgos.
- Los rasgos tienen un grado razonable de estabilidad.
- Nuestra cuantificación de los rasgos tiene relación con la conducta real.

5. Las tres preguntas fundamentales en el campo de las pruebas se relacionan con:

- Confiabilidad, es decir, la estabilidad de la medida.
- Validez, es decir, lo que una prueba en realidad mide.
- Normas, es decir, el marco para interpretar las puntuaciones de la prueba.

Estudiamos estos temas en profundidad en los capítulos 3, 4 y 5. El modo en que se elaboran las pruebas, tratado en el capítulo 6, y las preocupaciones prácticas, como el tiempo y costo, también son aspectos importantes a considerar.
6. Identificamos siete períodos principales en la historia de las pruebas. Comprender los temas predominantes en ellos ofrece una perspectiva sobre los temas actuales. Los períodos y los nombres que les dimos son:

- Hasta 1840 Antecedentes remotos
- 1840-1880 Creación del escenario
- 1880-1915 Raíces
- 1915-1940 Florecimiento
- 1940-1965 Consolidación
- 1965-2000 Pasado reciente
- 2000 al presente Actualidad

7. Identificamos seis fortalezas principales que influyeron en el desarrollo del campo de

las pruebas tal y como existe en la actualidad: el impulso científico, la preocupación por el individuo, las aplicaciones prácticas, la metodología estadística, el ascenso de la psicología clínica y las computadoras.

8. Desarrollamos la siguiente definición de seis elementos de una prueba: Una prueba es un proceso o dispositivo estandarizado que ofrece información sobre una muestra de la conducta o los procesos cognitivos de una manera cuantificada.

Palabras clave

aplicación adaptable por computadora
Binet, Alfred
calificación automatizada
Cattell, James McKeen
confiabilidad
ejecución máxima
ejecución típica
escáner
Galton, Francis
informes interpretativos
interpretación referida al criterio
interpretación referida a la norma
medidas de intereses vocacionales
normas
perspectiva diferencial
práctica basada en la evidencia
prueba de ejecución
prueba de lápiz y papel
prueba de poder
prueba de velocidad
prueba estandarizada
prueba grupal
prueba individual
pruebas de capacidad mental
pruebas de rendimiento
pruebas neuropsicológicas
pruebas objetivas de personalidad
Spearman, Charles
técnicas proyectivas
teoría clásica de las pruebas
teoría de la respuesta al reactivo
validez

Ejercicios

1. Por medio de la biblioteca de tu universidad, consulta el artículo de Cattell de 1890, donde acuñó el término "prueba mental". Busca en las referencias de este libro para encontrar los datos completos del artículo. O también puedes encontrarlo en http://psychclassics.yorku.ca. Revisa la lista de pruebas que ahí se describen. ¿Qué piensas de estas pruebas como predictores del éxito académico?
2. Si sucede que estás tomando un curso de historia al mismo tiempo que lees este libro, trata de relacionar algo de tu curso con los períodos del desarrollo en la historia de las pruebas. ¿Encuentras alguna tendencia o fortaleza en tu curso que pueda haber influido en el desarrollo de este campo?
3. La mayoría de las universidades, incluso muchos departamentos dentro de grandes universidades, tienen su propio escáner para procesar las hojas de respuesta de las pruebas. Intenta localizar un escáner para que veas cómo trabaja. ¿Cuál es el "producto" del escáner?
4. Piensa en las pruebas que has contestado aparte de las pruebas en el salón de clases. Clasifica cada una de ellas de acuerdo con estas distinciones: a) lápiz y papel o ejecución, b) velocidad o poder, c) de aplicación individual o grupal, d) de ejecución máxima o típica y e) referidas a la norma o al criterio para la interpretación.
5. Consulta la página http://psychclassics.yorku.ca/ para acceder al trabajo clásico de Alfred Binet *New Methods for the Diagnosis of the Intellectual Level of Subnormals* [Nuevos métodos para el diagnóstico del nivel intelectual de los subnormales], escrito en 1905. (Toma nota del uso de palabras como imbécil e idiota, que hoy son considerados términos peyorativos, pero que entonces eran descriptores clínicos estándar.) A partir de la lectura de los primeros párrafos de la obra de Binet, ¿qué crees que intentaba hacer?
6. Entra a la página http://www.nces.ed.gov/nationsreportcard para ver los resultados del National Assessment of Educational Progress (NAEP). ¿Cuántos grados evaluó NAEP? ¿De cuántas materias escolares hay informes disponibles? Accede a los informes de alguna materia que te interese. ¿Cuáles son algunos de los principales hallazgos sobre esa materia?
7. Aquí hay tres rasgos: altura, inteligencia, cordialidad. ¿En cuál de ellos crees que la gente difiere más?
8. Recuerda nuestro comentario de que muchas pruebas se conocen principalmente por sus iniciales. Ve si puedes recordar los nombres completos que corresponden a las siguientes iniciales.

GRE EDI SII
LSAT BDI-II

9. Muchos documentos clásicos de la historia de las pruebas (p. ej., los de Darwin, Galton, Cattell y Binet) pueden consultarse en esta página: http://psychclassics.yorku.ca/. Revísala. Échale un ojo a algunos documentos para

hacerte una idea de cómo se acercaban los autores a sus temas.

10. Para ver una presentación gráfica interesante de las relaciones entre la gente que trabajó en las primeras pruebas de inteligencia, revisa esta página: http://www.indiana.edu/~intell. Haz clic en Interactive Map. ¿En dónde entra Piaget? ¿Qué hay de Anastasi? Puedes acceder a biografías breves de la mayoría de los personajes que aparecen en nuestra historia del campo haciendo clic sobre su nombre en el mapa interactivo.

Notas

[1] En este capítulo, nos referimos sólo a la primera edición de las pruebas. En los siguientes capítulos, nos referimos a las ediciones más recientes y a sus iniciales correspondientes, por ejemplo, WAIS-IV, MMPI-2 y así sucesivamente.

[2] Por muchos años, esta prueba se llamó *Scholastic Aptitude Test* [Prueba de Aptitudes Escolares]. El título se cambió de manera oficial a *Scholastic Assessment Test* [Prueba de Evaluación Escolar] en 1992 y después simplemente a SAT. Estos antiguos nombres aún aparecen en muchas publicaciones. Aquí se refiere en particular al SAT I: Prueba de razonamiento. El SAT II: Prueba de materias es una serie de pruebas en áreas específicas como literatura, francés o química.

[3] Seguimos la práctica moderna de referirnos a ésta como la Escala Binet-Simon. En su propio trabajo, Binet no empleó un nombre oficial para esta prueba; simplemente se refirió a ella como la "escala" o la "prueba".

CAPÍTULO 2
Fuentes de información sobre las pruebas

Objetivos

1. Identificar las nueve fuentes principales de información sobre las pruebas.

2. Evaluar las fortalezas y debilidades de cada fuente.

3. Determinar dónde puedes encontrar una copia impresa de las fuentes en tu biblioteca.

4. Ser competente para acceder a las versiones electrónicas de las fuentes.

Dos problemas comunes que requieren información sobre las pruebas

Uno de los problemas prácticos más comunes en el campo de las pruebas psicológicas es encontrar información sobre una prueba en particular. El problema se presenta en dos contextos. Primero, cuando oímos acerca de una prueba, sabemos poco de ella y necesitamos información sobre sus características. Los siguientes son ejemplos de este problema.

- Un artículo de revista cita el *ABC Personality Inventory* [Inventario ABC de Personalidad]. No conoces esta prueba, pero quieres saber qué información está disponible sobre su validez.
- Eres psicólogo escolar. Recibes un informe de otra escuela sobre un estudiante que fue transferido a tu escuela. El informe incluye puntuaciones del *Pennsylvania Nonverbal Intelligence Test* [Prueba de Inteligencia No Verbal Pennsylvania]. Necesitas conocer las normas de esta prueba.
- Ves un anuncio en una revista profesional del *Vasquez Index of Life Stress* [Índice Vasquez de Estrés en la Vida]. Te preguntas si esta prueba ha sido reseñada en algún lugar.

El segundo problema requiere información por completo diferente. A veces necesitamos determinar qué pruebas existen para propósitos particulares. Los siguientes son ejemplos de este problema.

- Trabajas en el departamento de recursos humanos de una corporación importante y necesitas una prueba de destreza manual para seleccionar a cierto tipo de empleados. ¿Ya existen pruebas diseñadas para este propósito?
- Eres consejero escolar que colabora en una comisión de pruebas en la escuela y que busca una prueba diagnóstica de lectura. ¿Qué pruebas debe revisar la comisión para tomar una decisión?
- Tu investigación del desarrollo matemático en niños te lleva a pensar que la capacidad cuantitativa puede relacionarse con la capacidad de pensamiento creativo. ¿Qué prueba de pensamiento creativo debes considerar usar en tu proyecto de investigación?

Ambos problemas –encontrar información sobre una prueba en particular y obtener una lista de pruebas que podrían usarse con cierto propósito– son muy comunes. Además, los psicólogos encuentran los problemas cada vez con mayor frecuencia, ya que la cantidad de pruebas aumenta. Antes, se podía conocer de manera razonable una buena parte de las pruebas disponibles, pero eso ya no es posible. Cada año, aparecen docenas

de pruebas nuevas o revisadas; algunas de las bases de datos que examinamos en este capítulo contienen más de 10 000 pruebas. Por lo tanto, existe una gran necesidad de ser capaz de obtener información sobre ellas. Un informe de la *American Psychological Association* (Turner, DeMers, Fox, & Reed, 2001) subrayó la responsabilidad profesional de ser competente en la selección de pruebas. De hecho, este informe sobre los requerimientos del usuario de pruebas ubica "selección de las pruebas apropiadas" inmediatamente después de la consideración de la confiabilidad, validez y normas entre los conocimientos y las habilidades que el usuario de las pruebas debe poseer. El consejo de administración de la *Society for Personality Assessment* [Sociedad para la Evaluación de la Personalidad] (2006) emitió una declaración similar. La habilidad para elegir y evaluar pruebas ha adquirido mayor importancia a medida que el número de pruebas nuevas y revisadas ha aumentado en los años recientes.

Por fortuna, hay excelentes fuentes de información que ayudan a responder a estas preguntas. En este capítulo se describen las fuentes de información y se examinan sus fortalezas y defectos.

Resumen de puntos clave 2-1

Dos preguntas comunes sobre las pruebas

1. ¿Cómo obtengo información sobre una prueba en particular?

2. ¿Qué pruebas están disponibles para cierto propósito?

Materiales de una prueba

Una vez que se ha identificado una prueba específica, se debe examinar la prueba misma y su manual. La inspección inicial suele basarse en lo que se denomina **materiales** de la prueba (distintas casas editoriales emplean nombres diferentes para dichos materiales). Los materiales incluyen una copia de los elementos básicos de la prueba, entre los cuales suelen estar un cuadernillo y otros estímulos, las instrucciones de aplicación, documentos para registrar las respuestas y un manual técnico, como se muestra en la figura 2-1. Las fuentes de información que se tratan en este capítulo son **externas** a la prueba misma. En el apéndice A se dan sugerencias para llevar a cabo una revisión de materiales de pruebas reales.

Todos los capítulos restantes de este libro utilizan las fuentes que presentamos en este capítulo. En particular, se recomienda a menudo al lector consultar estas fuentes para encontrar ejemplos de los temas que se tratan en los otros capítulos. Por ejemplo, cuando discutamos las normas en el siguiente capítulo, se pedirá al lector encontrar ejemplos de normas; en el capítulo 8, al discutir las pruebas de inteligencia, se pedirá al lector encontrar información sobre estas pruebas en las fuentes que presentamos aquí. Por tanto, es importante ser competente al usar estas fuentes a partir de ahora.

Identificamos nueve fuentes principales de información sobre las pruebas: listas exhaustivas de pruebas, reseñas sistemáticas de pruebas publicadas, listados electrónicos, colecciones para propósitos especiales, libros sobre pruebas únicas, libros de texto sobre el campo de las pruebas, revistas profesionales, catálogos de las editoriales y otros usuarios de pruebas. En las siguientes secciones, describiremos cada una de estas fuentes de información; después, compararemos sus fortalezas y debilidades, sobre todo en relación con los dos problemas que identificamos más arriba.

Nota especial: Antes de leer las siguientes dos secciones, tal vez desees echar una mirada rápida a la tercera sección sobre los listados electrónicos (33a»). Mucho de lo que se trata en las siguientes dos secciones puede encontrarse en internet. Primero presentamos el tratamiento de las **copias impresas**, porque éstas antecedieron históricamente a las versiones electrónicas, y establecemos el formato básico de las versiones electrónicas.

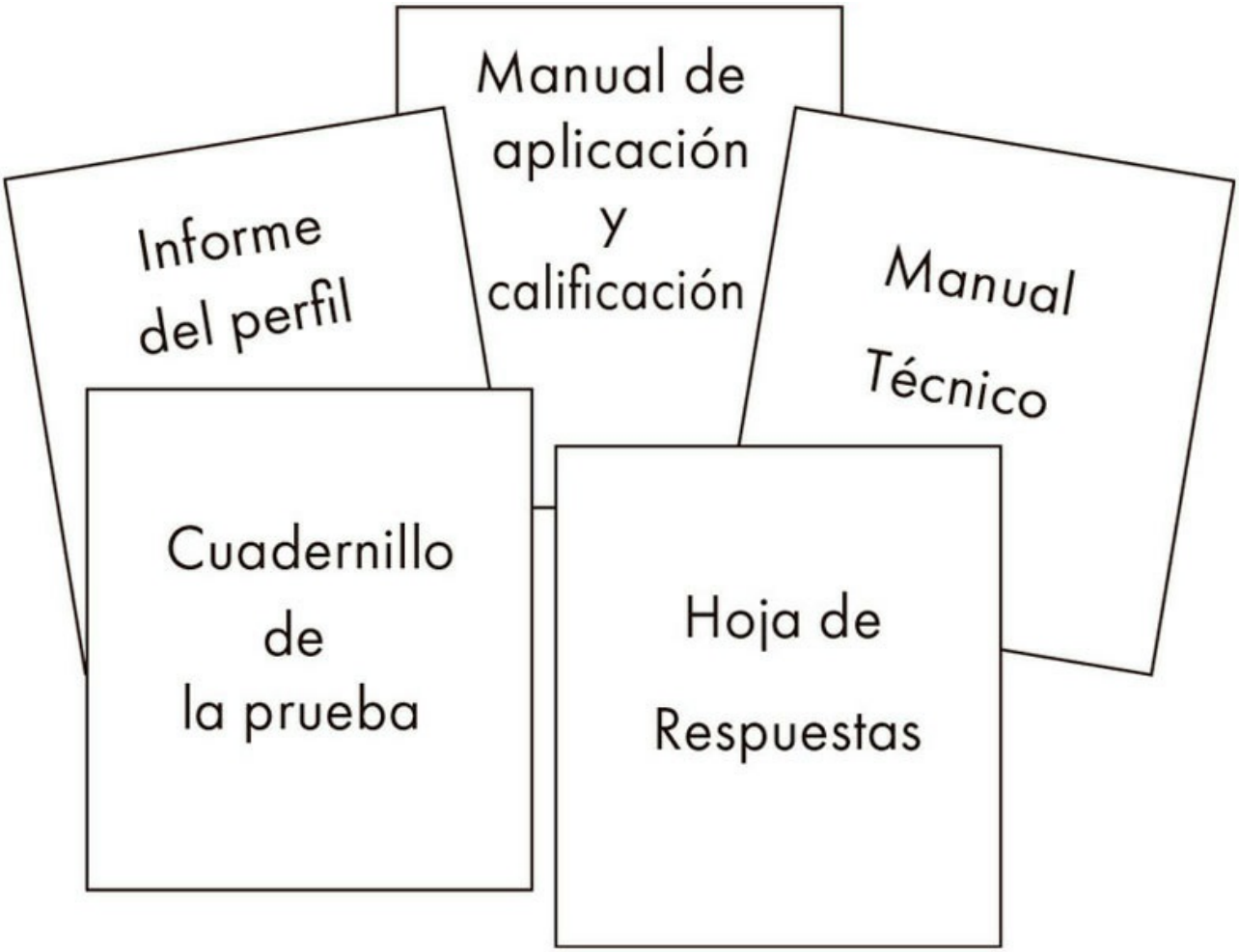


Figura 2-1. Contenidos típicos de los materiales de una prueba.

Resumen de puntos clave 2-2

Nueve fuentes principales de información sobre las pruebas
1. Listas exhaustivas de pruebas
2. Reseñas sistemáticas de pruebas publicadas
3. Listados electrónicos
4. Colecciones para propósitos especiales
5. Libros sobre pruebas únicas
6. Libros de texto sobre el campo de las pruebas
7. Revistas profesionales
8. Catálogos de las editoriales
9. Otros usuarios de pruebas

Listas exhaustivas de pruebas

Existen tres listas exhaustivas de pruebas, las cuales fueron diseñadas para ofrecer información básica descriptiva, por ejemplo, las fuentes para identificar el propósito de una prueba, si tiene subpruebas, tiempo de aplicación y casa editora. Estas fuentes **no** evalúan la calidad de la prueba.

Dos de estas listas sólo incluyen pruebas publicadas: *Tests in Print* [Pruebas Publicadas] y *Tests* [Pruebas]. La otra lista incluye sólo pruebas no publicadas: *The Directory of Unpublished Mental Measures* [Directorio de Medidas Mentales No Publicadas].

Tests in Print VIII (Murphy, Geisinger, Carlson, & Spies, 2011), popularmente conocido como TIP, intenta hacer una lista de todas las pruebas que se han publicado y comercializado en inglés. TIP contiene entradas de aproximadamente 3000 pruebas; cada nueva edición aparece en un lapso de 3 a 5 años. La primera salió a la luz en 1961 y estuvo a cargo de Oscar Krisen (O. K.) Buros, a quien podría llamarse el santo patrono de la información sobre las pruebas. Él también creó el *Mental Measurements Yearbook* [Anuario de Mediciones Mentales] (MMY), del que se habla en la siguiente sección. La figura 2-2 presenta una entrada del TIP como ejemplo. Examina la figura para determinar el tipo de información que proporciona. TIP también consigna todas las reseñas que aparecen en el MMY, que se describe más adelante. Sin embargo, no todas las pruebas que aparecen en el TIP se han reseñado en el MMY.

[573]

Clinical Assessment of Depression
[Evaluación Clínica de la Depresión]

Objetivo: "Diseñada como auxiliar en la evaluación clínica y diagnóstico de la depresión."
Población: De 8 a 79 años de edad.
Fechas de publicación: 1994-2004.
Acrónimo: CAD.
Puntuaciones, 14: Puntuaciones en las escalas de síntomas (Estado de ánimo deprimido, Ansiedad/Preocupación, Interés reducido, Fatiga cognitiva y física, Total), Puntuaciones de las escalas de validez (Inconsistencia, Impresión negativa, Infrecuencia), Puntuaciones críticas de grupos de reactivos (Desesperanza, Autodevaluación, Sueño/Fatiga, Fracaso, Preocupación, Nerviosismo).
Aplicación: Grupal.
Precio en 2007: 130 dólares por los materiales, que incluyen manual profesional (2004, 96 páginas), 25 hojas de calificación y 25 formatos para el perfil/resumen de las puntuaciones; 185 dólares por el programa de calificación (CD-ROM), que incluye calificaciones e informes ilimitados.
Tiempo: 10 minutos.
Autores: Bruce A. Bracken y Karen Howell.
Editorial: Psychological Assessment Resources, Inc.
Referencias cruzadas: Para consultar la reseña de Michael G. Kavan y Jody L. Kulstad, véase 17:48.

Figura 2-2. Entrada muestra de *Tests in Print.*
Fuente: Murphy, L. L., Geisinger, K. F., Carlson, J. F., & Spies, R. A. (Eds.). *Tests in print VIII*, p. 169. Copyright © 2011. Traducido del original por Editorial El Manual

Moderno con autorización de Buros Center for Testing.

La segunda lista es *Tests: A Comprehensive Reference for Assessments in Psychology, Education and Business* [Libro de Consulta Completo para las Evaluaciones en Psicología, Educación y Negocios], quinta edición (Maddox, 2003). Por lo general, se cita esta lista sólo como *Tests*; sus primeras ediciones, a cargo de D. J. Keyser y R. C. Sweetland, aparecieron en 1983, 1986 y 1991. Al igual que TIP, *Tests* incluye sólo medidas publicadas en inglés. La casa editora de esta lista, Pro-Ed, no tiene planeada una nueva edición en el futuro inmediato, por lo que esta fuente será menos útil.

La tercera lista es *Directory of Unpublished Experimental Mental Measures* (Goldman & Mitchell, 2008). Como lo sugiere su nombre, esta fuente, en contraste con TIP y *Tests*, se centra en medidas que no han sido publicadas; no incluye pruebas disponibles de editoriales comerciales. Publicado por la *American Psychological Association*, el *Directory*más reciente, que es el volumen 9, contiene 7605 entradas; también incluye listas acumulativas de pruebas que aparecieron en los volúmenes previos. Las entradas del volumen 9 se obtuvieron escaneando 34 revistas del periodo 2001-2005. La figura 2-3 presenta una entrada muestra.

10308
Nombre de la prueba: Mathematics Anxiety Scale For Children [Escala para Niños de Ansiedad ante las Matemáticas]
Objetivo: Identificar ansiedad ante las matemáticas en niños.
Número de reactivos: 22.
Formato: Las escalas de calificación van de 1 *(no nervioso)* a 4 *(muy, muy nervioso)*. Se aplican todos los reactivos.
Confiabilidad: Coeficiente alfa de .92.
Validez: Las correlaciones con otras variables variaron entre −.21 y .53.
Autor: Beasley, T. M. *et al.*
Artículo: A confirmatory factor analysis of the Mathematics Anxiety Scale for Children.
Revista: *Measurement and Evaluation in Counseling and Development*, abril, 2001, *34*, 14-26.
Investigación relacionada: Chiu, L. H., & Henry, L. L. (1996). Development and validation of the Mathematics Anxiety Scale for Children. *Measurement and Evaluation in Counseling and Development, 23*, 121-127.

Figura 2-3. Entrada muestra del *Directory of Unpublished Experimental Mental Measures*.
Fuente: Goldman, B. A., & Mitchell, D. F. *Directory of Unpublished Experimental Mental Measures* (Vol. 9, p. 23). Copyright © 2008 por la *American Psychological Association*. Reproducida con autorización. El uso de información de la APA no implica el respaldo de esta organización.

¡Inténtalo!
Revisa en tu biblioteca cuáles listas exhaustivas están disponibles. Haz un registro de la información de acceso para usarla en ejercicios posteriores.

Fuente	Ubicación
TIP	_____
Tests	_____
Directory of Unpublished...	_____

Reseñas sistemáticas

La segunda fuente de información principal consta de dos series continuas que ofrecen reseñas sistemáticas de pruebas publicadas. Las reseñas contienen juicios acerca de la calidad de la medición de las pruebas, así como descripciones extensas de sus propósitos y materiales. La "calidad de la medición" incluye características como confiabilidad, validez, normas y procedimientos de elaboración de la prueba, justo los temas de los siguientes cuatro capítulos de este libro.

La primera y más conocida fuente de reseñas sistemáticas es el *Mental Measurements Yearbook*, cuya más reciente edición es la número dieciocho (Spies, Carlson, & Geisinger, 2010). La primera edición apareció en 1938 y constó sólo de 83 páginas, a diferencia de las casi 1000 páginas de la edición actual. O. K. Buros fue el creador de esta serie, por lo que esta publicación, en cualquiera de sus ediciones, a menudo se cita simplemente como "**Buros**" o el "**MMY de Buros**". Buros supervisó la preparación de las primeras ocho ediciones, pero, después de su muerte en 1978, el *Buros Institute of Mental Measurements* [Instituto Buros de Medición Mental] de la Universidad de Nebraska-Lincoln llevó a cabo el trabajo. Cada dos o tres años se publica una nueva edición del MMY. Las reseñas, por lo general, son preparadas por un autor experto en el área de contenido que aborda la prueba o por un experto en pruebas. Por ejemplo, una prueba de lectura puede ser reseñada por un especialista en lectura y una prueba de intereses vocacionales, por un consejero escolar, o bien un profesor de medición educativa y psicológica puede reseñar cualquiera de los dos. La mayoría de las entradas contiene dos reseñas independientes de la prueba.

Las reseñas del MMY no siguen un formato rígido, pero suelen cubrir los siguientes aspectos. Primero, se identifica el objetivo de la prueba y, después, se hace una descripción de los materiales, como el número y el tipo de reactivos, límite de tiempo, modo de aplicación, número de niveles y formas, tipos de calificación y manuales disponibles. Luego, la reseña brinda información sobre la validez, confiabilidad y normas de la prueba. En estas discusiones, el autor de la reseña ofrece comentarios críticos sobre lo adecuado de estas características técnicas. La reseña, a menudo, concluye con recomendaciones sobre los posibles usos de la prueba o con recomendaciones de que no se use.

Los MMY recientes tienen seis índices útiles. El índice de títulos de pruebas es una lista ordenada de manera alfabética de los títulos oficiales de las pruebas. El índice de acrónimos ofrece una lista de las abreviaturas comunes de las pruebas (p. ej., GRE de *Graduate Record Examination* [Examen de Registro de Graduados]). El índice de temas clasificados presenta listas de pruebas agrupadas en amplias categorías, por ejemplo, lectura o bellas artes. El índice y directorio de editoriales incluye nombres y direcciones de todas las editoriales cuyas pruebas están reseñadas. El índice de nombres presenta a todos los autores de pruebas y reseñas, y a quienes están en las referencias que se incluyen en el volumen. Por último, el índice de puntuaciones ofrece una lista de todas

las puntuaciones que se obtienen de las pruebas reseñadas en el volumen.

La segunda fuente es ***Test Critiques*** [Críticas de Pruebas], cuyo más reciente volumen es el XI (Keyser, 2004). Creado por Keyser y Sweetland en 1984, *Test Critiques* es publicado por Pro-Ed, la misma organización que publica *Tests*, antes descrito. Las reseñas de *Test Critiques* son muy parecidas en alcance y carácter a las del MMY; sin embargo, *Test Critiques* no cubre tantas pruebas como el MMY. Igual que éste, *Test Critiques* ofrece distintos índices que son de utilidad, como el índice de títulos de pruebas y el de editoriales, además de un índice acumulativo de reseñas de todos los volúmenes de la serie. La editorial de *Test Critiques*, Pro-Ed, no tiene planeado uno nuevo en el futuro inmediato, así que esta fuente será menos útil.

Resumen de puntos clave 2-3

Dos fuentes principales de reseñas de pruebas

1. El Mental Measurements Yearbook (MMY) de Buros
2. Test Critiques

Listados electrónicos «33a

Hay cuatro fuentes electrónicas importantes de información acerca de las pruebas. Usar estas fuentes puede aumentar en gran medida la eficacia al buscar información.

Colección de pruebas ETS en la red

Por muchos años, el *Educational Testing Service* [Servicio de Pruebas Educativas] ha mantenido un inventario de pruebas. En el curso de los años, existió en distintas formas, incluyendo una colección especial de libros y microfichas, pero en la actualidad, la forma más usual de la colección es accesible en internet, en la dirección http://www.ets.org/test_link/about. Si esta dirección cambia, se puede acceder a la colección buscando en internet **ETS Test Collection** o ETS Test Link. Por ahora, la base de datos tiene más de 25 000 entradas.

Las pruebas que aparecen en la base de datos ETS *Test Collection* pueden buscarse por título, autor, iniciales usadas comúnmente (acrónimos) o palabras clave (es decir, nombres de variables o temas como ansiedad o creatividad), así como por otros medios. Las entradas en *Test Collection* son muy parecidas a las de TIP y *Tests*; es decir, ofrecen información descriptiva básica acerca de las pruebas. Quizá el uso más común de *Test Collection* es para identificar pruebas disponibles para cierta variable. Una búsqueda sobre un "tema" como la depresión arrojará una lista de 20 o 30 pruebas. También es de especial utilidad cuando se tiene sólo el nombre de una prueba (o incluso sólo sus iniciales) y se necesita saber el nombre de la editorial o del autor.

Reseñas de Buros disponibles en versión electrónica

Las reseñas en el *Mental Measurements Yearbook* de Buros están disponibles en versión electrónica desde la décima edición. Se pueden consultar de dos maneras; primero, muchas bibliotecas académicas se suscriben a un producto de EBSCOhost que contiene copias de las reseñas íntegras. Este producto permite hacer búsquedas muy parecidas a las de bases de datos como PsycINFO; es decir, se introduce el nombre de la prueba y, si ya ha sido reseñada, las reseñas aparecerán. Si una biblioteca se suscribe a este producto, se puede acceder a una reseña de Buros sin cargo adicional. Pero hay un truco en las últimas reseñas en línea: ¡puedes hacer clic en un botón y escuchar una reseña (en uno de tres diferentes acentos del inglés)!

El *Buros Center for Testing* (http://buros.org/) también permite adquirir en línea reseñas de pruebas, pero se requiere de pago por el uso de este servicio (actualmente es de 15 dólares por reseña). Un beneficio adicional interesante del sitio BIMM es que permite buscar palabras clave de las pruebas en el inventario de reseñas. Por ejemplo, se puede buscar mediante la palabra clave "creatividad" y obtener una lista de pruebas

relacionadas con este constructo, además de una indicación sobre si existe o no una reseña en Buros de dichas pruebas.

Resumen de puntos clave 2-4

Dos sitios importantes en internet para buscar pruebas
Para entrar al ETS Test Collection:
http://www.ets.org/test_link/about
Para entrar al Buros Center for Testing:
http://buros.org/

¡Inténtalo!

Entra en dos páginas web mencionadas en esta sección. Si es posible, márcalas como favoritas para facilitar el acceso en ejercicios futuros. También verifica si tu biblioteca tiene la base de datos de reseñas MMY. Pregunta al bibliotecario o ve el listado de la biblioteca de bases de datos electrónicas.

Health and Psychosocial Instruments (HaPI)

Health and Psychosocial Instruments [Instrumentos de Salud y Psicosociales], que se abrevia HaPI, es una base de datos en que se describen más de 15 000 pruebas, escalas de valoración, cuestionarios, etc., obtenidos principalmente de artículos de revistas especializadas en salud y ciencias psicosociales. Esta base de datos es elaborada por *Behavioral Measurements Database Services* [Servicios de Bases de Datos de Mediciones Conductuales] (BMDS). Está disponible en CDROM en BMDS o en línea en Ovid y EBSCO. La base de datos HaPI no cuenta con versión impresa.

PsycTESTS

El nuevo chico en la cuadra es **PsycTESTS**, producto de la *American Psychological Association*. Lanzado en 2011, PsycTESTS es una base de datos que contiene información acerca de pruebas y, en algunos casos, las pruebas mismas. Se concentra en pruebas "no publicadas", es decir, las que no están disponibles para su compra, pero aparecen en alguna fuente publicada, como un artículo de revista. El sitio de internet de PsycTESTS (http://www.apa.org/pubs/databases/psyctests/index.aspx) dice que también incluye pruebas disponibles comercialmente y las ligas con las editoriales. Como con otras bases de datos de la APA (p. ej., PsycINFO y PsycARTICLES), sólo se puede tener acceso a PsycTEST por medio de una suscripción.

Desde 2012, PsycTESTS contiene cerca de 10 000 pruebas; es muy pronto para decir si este producto "tendrá éxito" entre los usuarios que buscan información sobre pruebas. Seguiremos pendientes.

Colecciones para propósitos especiales «34-35a

Las colecciones para propósitos especiales brindan información sobre pruebas en un rango de temas limitado; en ellas, la información es, por lo común, como la de los listados exhaustivos antes descritos, es decir, básica y descriptiva. Sin embargo, en algunos casos, estas fuentes incluyen la prueba completa y comentarios breves de evaluación. Las colecciones para propósitos especiales, a menudo, incluyen instrumentos publicados y no publicados; también es muy común que incluyan una revisión general de aspectos pertinentes para el área a la que pertenecen las medidas.

Al aumentar la cantidad y variedad de pruebas, estas colecciones han sido un complemento valioso de los listados exhaustivos, pero más incómodos, descritos con anterioridad. Todo el tiempo aparecen nuevas colecciones para propósitos especiales; la siguiente es una lista representativa, aunque no exhaustiva, de ellas:

- El *Handbook of Personality Assessment* [Manual de Evaluación de la Personalidad] (Weiner & Greene, 2008) ofrece capítulos separados para nueve de las medidas de personalidad de uso más generalizado; incluye historia, interpretación práctica y características psicométricas de cada una.
- En una línea parecida, el *Integrative Assessment of Adult Personality* [Evaluación Integral de la Personalidad Adulta] (Harwood, Beutler, & Groth-Marnat, 2011) dedica capítulos separados a cado uno de los instrumentos clínicos más populares y sus aplicaciones en diversos escenarios.
- El *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health* [Manual de Escalas de Valoración Clínica y Evaluación en Psiquiatría y Salud Mental] (Baer & Blais, 2010) dedica capítulos separados a escalas para medir ansiedad, trastorno bipolar, adicción al alcohol y a la nicotina, trastornos alimentarios y muchos otros síndromes. Cada capítulo enumera escalas (pruebas) para su área y ofrece una sinopsis de la información sobre confiabilidad, validez y normas de cada una.
- El *Handbook of Assessment Methods for Eating Behaviors and Weight-Related Problems: Measures, Theory, and Research* [Manual de métodos de evaluación para las conductas alimentarias y problemas relacionados con el peso: medida, teoría e investigación] (Allison & Baskin, 2009) ofrece una colección exhaustiva de métodos de evaluación relacionados con todos los aspectos de los trastornos alimentarios, incluyendo, por ejemplo, medidas de imagen corporal, ingesta de alimentos e, incluso, actitudes hacia las personas obesas.
- *Measures for Clinical Practice and Research: A Sourcebook* [Medidas para la práctica clínica y la investigación: libro de fuentes](Fischer & Corcoran, 2007) es una colección de dos volúmenes de medidas de lápiz y papel simples para

constructos clínicamente pertinentes. En el primer volumen aparecen medidas para familias, niños y parejas, mientras que el segundo contiene medidas para otros tipos de adultos.

- *Scales for the Measurement of Attitudes* [Escalas para la Medición de Actitudes] (Shaw & Wright, 1967] también es una excelente, aunque un poco anticuada, colección de escalas para medir actitudes. Se trata de medidas para una serie inimaginable de actitudes. Se incluyen las escalas junto con breves sumarios de confiabilidad, validez y normas de cada una.
- *A Counselor's Guide to Career Assessment Instruments* [Guía Guía de instrumentos del orientador vocacional para la elección de carreras] de Whitfield *et al*. (2008), ahora en su novena edición, ofrece una cobertura excelente y actualizada de muchos instrumentos pertinentes para los procesos de asesoría sobre elección de carrera.
- Por último, mencionamos *Positive Psychological Assessment: A Handbook of Models and Measures* [Evaluación en Psicología Positiva: Manual de Modelos y Medidas] de Lopez y Snyder (2003), que considera varias medidas para una gran cantidad de constructos de la psicología positiva, como gratitud, empatía y sentido del humor. De este modo, constituye una base de mediciones para el área de la psicología positiva, que se encuentra en pleno auge.

Todos estos libros ilustran los intentos de catalogar y evaluar los instrumentos de medición en las áreas elegidas. Son fuentes invaluables tanto para la práctica como para la investigación.

Libros sobre pruebas específicas «35a

Algunas pruebas, aunque muy pocas, tienen un uso tan generalizado que son tema de libros enteros, en los cuales, por lo general, se presentan estudios de caso basados en dichas pruebas, se describen los perfiles de puntuaciones en ciertos casos o se exploran maneras especiales de calificarlas. Ejemplos de esto son los libros sobre las escalas de inteligencia Wechsler (véase capítulo 8), el Inventario Multifásico de Personalidad de Minnesota (MMPI; véase capítulo 13) y la prueba Rorschach de manchas de tinta (véase capítulo 14).

- Las escalas Wechsler han aparecido en numerosos libros; un buen ejemplo es el que editaron Prifitera, Saklofske y Weiss (2008), que incluye consejos de expertos para usar e interpretar el WISC-IV. Incluye capítulos acerca de grupos de niños con problemas de aprendizaje o con superdotación intelectual. Hay un trabajo similar sobre el WAIS-IV (Weiss, Saklofske, Coalson, & Raiford, 2010) que incluye capítulos sobre temas como el empleo de WAIS para la evaluación neuropsicológica y de la psicopatología.

- La obra en dos volúmenes de Dahlstrom, Welsh y Dahlstrom (1960, 1972) sobre el MMPI ayudó a sentar los estándares para los libros dedicados a una sola prueba. La obra ofrece numerosos estudios de caso con diversos perfiles del MMPI. Una gran cantidad de libros más recientes tratan del MMPI-2. Por ejemplo, Greene (2010) y Graham (2011) presentan un resumen completo de la investigación sobre el MMPI-2, así como extensas sugerencias para su interpretación.
- El Rorschach ha sido tema de muchos libros en los que se presentan estudios de caso y sugerencias para codificar las respuestas. Una serie de libros de Exner y colegas (p. ej., Exner, 2003), como se detalla con mayor profundidad en el capítulo 14, se ocupa de manera extensa de la calificación e interpretación de esta prueba. *Principles of Rorschach Interpretation* [Principios de la interpretación del Rorschach] de Weiner (2003) es otro buen ejemplo.
- El uso clínico del Test de Apercepción Temática (TAT) y sus derivados es tratado en detalle por Bellak y Abrams (1997) en un libro que ahora está en su sexta edición. Para muchos propósitos, este libro funciona como manual de uso del TAT. Jenkins (2008) hizo descripciones de numerosos sistemas para calificar el TAT.
- Por último, destacamos una serie de libros con títulos que comienzan "Essentials of..." que John Wiley & Sons ha publicado sobre las pruebas más populares, como WAIS, MMPI, los inventarios de Millon y el Rorschach. Esta serie cuenta ahora con más de dos docenas de libros. En el sitio de Wiley.com, puedes buscar "essentials of" para ver si hay un libro disponible que sea de tu interés.

Estos libros sobre pruebas específicas pueden ser fuentes ricas de información para aquellos psicólogos que usen mucho estas pruebas. Los estudios de caso pueden ser fascinantes; sin embargo, los libros, por lo general, suponen que el lector ya tiene un entrenamiento considerable en teoría y uso de pruebas. Estos libros no fueron diseñados para novatos; quizá más importante, hay muy pocas pruebas de las que se disponga un tratamiento tan amplio.

Libros de texto sobre el campo de las pruebas

Los libros de texto sobre el campo de las pruebas, como éste, no pretenden ser fuentes importantes de información sobre pruebas particulares, sino que buscan enseñar conceptos fundamentales sobre él (confiabilidad, validez, etc.) y ejemplifican los tipos de pruebas que están disponibles en varios dominios. Sin embargo, por realizar esta última tarea, los libros de texto son una fuente potencialmente útil de información sobre las pruebas, pues casi todos incluyen algunos ejemplos de pruebas de inteligencia y de rendimiento, inventarios objetivos de personalidad, técnicas proyectivas, entre otros. A menudo, aunque no siempre, los ejemplos son elegidos porque ilustran un enfoque inusual. Los autores de libros de texto, a veces, incluyen comentarios críticos sobre las pruebas que se usan como ejemplos.

Revistas profesionales «36-37a «36-37b

Muchas revistas especializadas de ciencias sociales y conductuales incluyen artículos relacionados con el uso de pruebas; sin embargo, ciertas revistas, con frecuencia, incluyen artículos en los que aparecen pruebas específicas o tratan los aspectos técnicos del campo de las pruebas. La mayoría de estos artículos no es sobre una prueba en particular, sino sobre sus características técnicas, como la confiabilidad o el análisis de reactivos. Los artículos son como temas avanzados de los siguientes cuatro capítulos de este libro. Sin embargo, estas revistas a veces contienen reseñas de pruebas y también artículos sobre las cualidades técnicas de pruebas específicas, en especial cuando se usan con poblaciones particulares.

Las siguientes revistas se concentran de manera casi exclusiva en pruebas y medición:

1. *Psychological Assessment* [Evaluación Psicológica]. Publicación trimestral, y una de las revistas oficiales de la *American Psychological Association*, que incluye artículos sobre una amplia variedad de temas relacionados con las pruebas, sobre todo en el terreno de la personalidad y la inteligencia.

2. *Journal of Personality Assessment* [Revista de Evaluación de la Personalidad]. Esta publicación bimestral es la revista oficial de la *Society for Personality Assessment*. Como lo sugiere el nombre de la revista, se concentra en pruebas relacionadas con la personalidad.

3. *Educational and Psychological Measurement* [Medición Educativa y Psicológica]. Revista clásica que cubre una amplia variedad de temas relacionados con las pruebas. La sección de validez en cada número está dedicada por completo a estudios sobre la validez de pruebas específicas.

4. *Journal of Educational Measurement* [Revista de Medición Educativa]. Revista de publicación trimestral del *National Council on Measurement in Education* [Consejo Nacional para la Medición en Educación] (NCME) que se concentra en asuntos técnicos relacionados con la medición en el campo de la educación. A veces incluye reseñas de pruebas.

5. *Measurement and Evaluation in Counseling and Development* [Medición y Evaluación en Orientación y Desarrollo]. Esta revista se ocupa del uso de las pruebas de particular pertinencia para la consejería pscológica. A menudo incluye reseñas de pruebas particulares.

6. *Applied Measurement in Education* [Medición Aplicada a la Educación]. De publicación trimestral, esta revista incluye artículos muy similares en carácter a los de JEM (véase número 4).

7. *Psychometrika*. Revista clásica sobre temas avanzados de estadística relacionados con la teoría y práctica de la medición. No es para los novatos en este campo.

8. *Educational Measurement: Issues and Practice* [Medición Educativa: Temas y Práctica]. Publicación trimestral del NCME (véase número 4) que presenta artículos que hacen hincapié en cuestiones prácticas de la medición educativa.

9. *Applied Psychological Measurement* [Medición Psicológica Aplicada]. Al igual que *Psychometrika*, esta revista trata temas avanzados de estadística y metodología relacionados con el campo de las pruebas. Incluye secciones especiales llamadas *Computer Program Exchange* [Intercambio de Programas de Cómputo] y Computer Software Reviews [Reseñas de Programas de Cómputo].

Las siguientes revistas se enfocan en áreas sustanciales de la psicología, pero, a menudo, incluyen artículos en que aparecen pruebas específicas o conceptos generales relacionados con el campo de las pruebas.

10. *Journal of Applied Psychology* [Revista de Psicología Aplicada]. Publicación bimestral de la *American Psychological Association* que se concentra en las aplicaciones en el área de recursos humanos e industriales/psicología organizacional, como selección de personal, medición del desempeño, capacitación o motivación para el trabajo.

11. *Personnel Psychology* [Psicología del Personal]. Esta publicación trimestral cubre áreas similares a las del *Journal of Applied Psychology*.

12. *Journal of Consulting and Clinical Psychology* [Revista de Orientación y Psicología Clínica]. De publicación bimestral, esta revista oficial de la *American Psychological Association* se ocupa, obviamente, de cuestiones clínicas, muchas de las cuales implican el uso de pruebas.

13. *Journal of Counseling and Development* [Revista de Orientación y Desarrollo]. De publicación trimestral, esta revista de la *American Counseling Association* [Asociación Americana de Orientación] incluye muchos artículos sobre el uso de pruebas en contextos de orientación psicológica.

14. *Journal of Clinical and Experimental Neuropsychology* [Revista de Neuropsicología Clínica y Experimental]. Esta revista se ocupa de la investigación de enfermedades, trastornos y disfunciones cerebrales; a menudo se enfoca en el papel de las pruebas en estas investigaciones.

15. *Intelligence*. Esta revista ofrece muchos ejemplos del uso de las pruebas para medir la inteligencia, entendida ampliamente.

Catálogos y personal de las editoriales

Una fuente clave de información sobre ciertos aspectos de una prueba es el catálogo de la editorial. Es evidente que esta fuente es útil sólo en el caso de las pruebas publicadas para su comercialización. El catálogo es fácilmente accesible si se lo solicita a la editorial y casi siempre está disponible en línea (véase el apéndice C para conocer los sitios de internet de las editoriales). En el caso de algunas de las pruebas más complejas, el catálogo puede contener 10 páginas de material. El catálogo de la editorial es la fuente primaria de información sobre cuestiones como el costo actual de la prueba, la variedad de hojas de respuesta y servicios de calificación disponibles, la edición más reciente de la prueba y otras cuestiones prácticas.

Las principales editoriales de pruebas tienen personal en sus oficinas centrales y representantes externos que cubre distintas áreas geográficas. Estos individuos pueden ser fuentes excelentes de información acerca de numerosos asuntos prácticos relacionados con las pruebas. Se pueden contactar usando los índices de la editorial en cualquiera de las listas integrales de pruebas presentadas anteriormente en este capítulo. Los sitios web de las editoriales también facilitan contactar a los representantes apropiados.

¡Inténtalo!

Revisa el sitio web de una de las editoriales enumeradas que aparecen en el apéndice C. Entra al sitio y determina si el catálogo completo de la editorial está disponible. También revisa junto con tu departamento académico si algún catálogo está disponible en versión impresa. Trata de ubicar, a partir del sitio web o del catálogo, a una persona que puedas contactar para pedir información sobre las pruebas de la editorial.

Otros usuarios de pruebas

Otros usuarios de pruebas pueden ser de utilidad como fuentes de información sobre las pruebas que usan actualmente. A menudo, sabrán sobre pruebas similares; por ejemplo, un psicólogo escolar sabrá qué pruebas se usan mucho para diagnosticar problemas de aprendizaje. Los consejeros en admisión a la universidad sabrán qué pruebas se usan para admitir alumnos. Los usuarios actuales de pruebas en un dominio particular pueden ser útiles para elaborar con rapidez una lista de pruebas posibles para propósitos específicos; desde luego, esta selección de pruebas no es un mero concurso de popularidad, sino que se necesita realizar un análisis crítico. Sin embargo, la tarea inicial es, a menudo, crear una lista de pruebas que, luego, deberán ser sometidas a un análisis crítico.

Los usuarios actuales de pruebas también pueden ser una fuente valiosa de información sobre características peculiares de una prueba que no son reveladas ni siquiera por el examen técnico más meticulosos posible, sino por el uso real de la prueba. Por ejemplo, el tiempo límite para una subprueba en una batería de pruebas múltiples puede ser demasiado estricto o demasiado generoso. Los examinados pueden tender a entender mal o a pasar por alto las instrucciones de una parte de la prueba, o las normas pueden estar gravemente sesgadas. Todas estas cuestiones prácticas pueden hacerse evidentes sólo para el usuario regular de una prueba.

Fortalezas y defectos de las fuentes

Cada fuente de información arriba descrita tiene fortalezas y defectos. El uso inteligente y competente de ellas requiere reconocer estos pros y contras, que se describen a continuación.

Las **listas exhaustivas** de pruebas son las más eficaces para lanzar una red grande y ver qué pruebas pueden estar disponibles para un propósito particular o para proporcionar información inicial sobre una prueba específica; por ejemplo, el rango de edad para el que fue diseñada la prueba, la editorial, etc. Estas listas no ofrecen evaluaciones críticas de la calidad de la prueba, de modo que una que cuenta con una extensa información sobre su validez y normas nacionales excelentes se encuentra en la misma lista que otra que prácticamente no cuenta con dicha información ni con normas.

Las **reseñas sistemáticas** son una fuente excelente de evaluaciones críticas de pruebas, pues ofrecen un servicio profesional sobresaliente. Sin embargo, no son perfectas. Primero, las reseñas de muchas pruebas simplemente no están disponibles. Segundo, la naturaleza del proceso de reseñar es tal que, a menudo, todavía no se encuentran en el caso de la edición más reciente de pruebas muy usadas y revisadas con frecuencia. Siempre hay un retraso, que a menudo es considerable, entre la publicación de una prueba y la aparición de sus reseñas. Mientras tanto, el potencial usuario necesita tomar una decisión acerca de usar o no la prueba. Tercero, la opinión del autor de una reseña es sólo eso: una opinión. El potencial usuario debe tomar en cuenta esa opinión, pero no es la última palabra sobre el valor de la prueba.

Los listados electrónicos, como el ETS Test Collection, ofrecen enormes mejoras en cuanto a accesibilidad de la información. Lo que podría tomar varias horas de búsqueda en fuentes impresas puede realizarse en minutos con estos sitios web. Es evidente que se necesita tener acceso en línea para aprovechar estos servicios; sin embargo, dicho acceso es de rutina en la actualidad. La limitación más importante es la misma que en los listados exhaustivos. Algunos de los listados electrónicos ofrecen sólo información descriptiva básica. Una limitación más importante es que las entradas pueden ser por completo obsoletas: una vez que entran en la base de datos, ahí se quedan para siempre. Así, el usuario necesita estar atento a lo anticuado e inaccesible de una entrada.

Las **colecciones para propósitos especiales** son muy útiles porque abarcan las pruebas disponibles en cierta área. Al mismo tiempo que se publican, tales colecciones están, por lo general, actualizadas. Sin embargo, no se actualizan con regularidad, de modo que en pocos años pueden volverse anticuadas por las nuevas pruebas o ediciones que se publican.

Los **libros sobre una sola prueba** suelen ser fuentes ricas en información; a menudo contienen estudios de caso interesantes, investigación más allá de lo que aparece en el manual de la prueba y fundamentos teóricos vastos para la interpretación de la prueba. Esta fuente tiene dos inconvenientes principales. Primero, estos libros están disponibles sólo para un reducido número de pruebas. Por lo general, se trata de las pruebas más

usadas; en cambio, es raro encontrar un libro entero dedicado a pruebas utilizadas en menor medida. Segundo, los libros tienden a estar escritos por personas que tienen un sentimiento muy positivo por la prueba, lo cual es comprensible, pero a veces conduce a hacer una evaluación demasiado optimista de la prueba.

Los **libros de texto** ofrecen un índice de la popularidad de una prueba. Al consultar varios de estos libros, se puede tener idea de qué pruebas pueden usarse mucho con propósitos específicos. Sin embargo, el hecho de que una prueba esté incluida en un libro de texto no debería tomarse como respaldo de la prueba, ni siquiera como un indicador válido de la frecuencia de su uso. Una prueba puede incluirse en un libro de texto porque ayuda a ejemplificar un enfoque particular de la construcción de pruebas y no porque es especialmente buena. Muchos libros de texto también tienen la costumbre de incluir pruebas demasiado anticuadas.

Las **revistas** enumeradas más arriba son fuentes de información muy importantes en relación con la investigación más reciente sobre las pruebas. Sin embargo, tienden a concentrarse más en el desarrollo metodológico de las pruebas que en las pruebas mismas. Cuando un artículo trata de alguna prueba específica, casi siempre el tema es algún uso especial de la prueba más que su uso más frecuente. Por ejemplo, un artículo puede investigar el uso de una subprueba de la Escala Wechsler de Inteligencia para Adultos con personas con deficiencias auditivas de más de 70 años de edad. Otro artículo puede investigar si las puntuaciones de una prueba son afectadas por un cambio en los límites de tiempo. Estas investigaciones se suman a los fondos de conocimiento acerca de una prueba, pero puede no ser útil para su uso ordinario.

Los **catálogos de las editoriales** son la mejor fuente de información sobre cuestiones prácticas como el costo de la prueba, el arreglo de los materiales disponibles para ella, los servicios de calificación, etc., pero no es una fuente útil para saber de su calidad. Los catálogos están llenos de frases como “una prueba confiable de...”, “medida validada de...”, “normas nacionales precisas”, entre otras. Estas frases no pueden tomarse al pie de la letra, pues, al fin y al cabo, la editorial está en el negocio de vender pruebas. Lo mismo puede decirse de los representantes de la editorial; ellos pueden ser fuentes esenciales de información sobre algunas cuestiones, pero sobre otras deben tomarse con reservas sus declaraciones.

Otros usuarios de pruebas pueden ser de gran utilidad para identificar qué pruebas se usan, por lo general, para un determinado propósito. También son fuentes importantes de información sobre las peculiaridades de la prueba; sin embargo, algunos de ellos no están actualizados con los últimos desarrollos en el campo. Las pruebas que utilizan puede estar determinado más por la inercia que por el juicio crítico y el conocimiento actualizado.

Todas las fuentes de información pueden ser útiles para algunos propósitos, pero no para otros. En última instancia, el usuario de la prueba debe ser capaz de combinar todas estas fuentes con su propia experiencia para juzgar la idoneidad de una prueba para propósitos particulares.

Resumen

1. Hay dos problemas comunes que requieren el uso de varias fuentes de información acerca de las pruebas.
2. Hay tres listas exhaustivas que ofrecen breves descripciones de un gran número de pruebas.
3. Hay dos fuentes principales de reseñas sistemáticas de evaluación de pruebas: Buros's MMY y Test Critiques.
4. Diversos sitios web proporcionan acceso electrónico a listados exhaustivos y reseñas sistemáticas.
5. Colecciones para propósitos especiales ofrecen listas de pruebas, a menudo con comentarios de evaluación sobre pruebas en campos específicos.
6. En el caso de algunas pruebas de uso generalizado, existen libros enteros dedicados a dar detalles sobre sus características e interpretación.
7. Los libros de texto sobre las pruebas ofrecen a menudo listas útiles para algunas pruebas muy empleadas.
8. Numerosas revistas profesionales ofrecen investigaciones sobre las pruebas; algunas también proporcionas reseñas de ellas.
9. Los catálogos de las editoriales son una fuente clave de información sobre cuestiones como el costo, la disponibilidad de ediciones nuevas o especiales y los servicios de calificación.
10. Los usuarios regulares de pruebas pueden ser fuentes de información útiles.
Todas las fuentes tienen sus fortalezas y debilidades particulares.

Palabras clave

Buros
equipo de muestra
ETS *Test Collection*
materiales de la prueba
MMY
PsycTESTS
Test Critiques
Tests
TIP

Ejercicios

Nota: Muchos de estos ejercicios pueden dividirse entre los miembros de una clase, y los resultados pueden compartirse entre todos.

1. Entra a uno de los sitios web enumerados en el Resumen de puntos clave 2-4

a. Introduce palabras clave y obtén una lista de pruebas. Aquí hay algunas palabras clave de muestra: pensamiento crítico, autoconcepto, lectura diagnóstica, agresión, depresión. O introduce tus propias palabras clave.

b. Toma el nombre de una prueba de los capítulos 8 al 15. Introdúcelo y ve qué información obtienes.

2. Visita tu biblioteca.

a. Encuentra cuáles listas exhaustivas de pruebas y cuáles de reseñas sistemáticas tiene la biblioteca. (Estos libros, por lo general, estarán en la sección de consulta general de la biblioteca). Toma nota de su ubicación para usarlas en ejercicios posteriores.

Buros MMY Ubicación: _____
Test Critiques Ubicación: _____
Tests in Print Ubicación: _____
Tests Ubicación: _____

b. Suponiendo que encontraras una edición del MMY, lee al menos dos reseñas. Elige las pruebas de tu interés. Nota la extensión y el estilo de las reseñas.

c. Suponiendo que encontraras una lista exhaustiva (TIP, *Tests*, etc.), observa la información que aparece en distintas entradas. Luego, toma un tema como el autoconcepto o el pensamiento crítico y usa el índice temático para encontrar las pruebas de ese tema.

3. Suponiendo que tienes acceso a la base electrónica de datos del *Mental Measurements Yearbook*:

a. Realiza una búsqueda de un título específico y entra al texto completo de una reseña.

b. Realiza una búsqueda de palabras clave, por ejemplo, autoconcepto, trastorno por déficit de atención o aprovechamiento en matemáticas. Luego, examina los registros que encuentres.

4. De cada una de las siguientes fuentes de información, revisa un libro o revista, sin leerlos detenidamente, y toma notas de qué tipo de información contienen:

a. Colecciones para propósitos especiales, que mencionamos en las páginas 34-35a».

b. Libros sobre pruebas específicas, que mencionamos en la página 35a».

c. Revistas, que mencionamos en las páginas 36-37a».

5. Usa las direcciones web de las editoriales que aparecen en el apéndice C y trata de entrar a sus catálogos. ¿Aparece el catálogo completo? Si es así, ¿qué información proporciona?

6. Selecciona **tres** revistas de las que se mencionan en las pp. 36-37b» e investiga si

tu biblioteca tiene suscripciones a ellas o acceso en línea al texto completo. Echa un ojo a los números recientes de las revistas para ver qué temas y pruebas tratan. Por favor, nota que algunos artículos serán muy técnicos y puedes no entender nada hasta que hayas leído los capítulos 3 al 6.

Nota: Los siguiente dos ejercicios son, en realidad, más ambiciosos y pueden ser adecuados como proyectos para todo un curso, pues, a diferencia de los ejercicios previos, no pueden realizarse en unos pocos minutos.

7. Usa el formato sugerido en el apéndice A, Reseña y selección de pruebas, para realizar una reseña formal de una prueba. Para este ejercicio, necesitarás los materiales de una prueba. También será útil contar con la entrada de esa prueba que aparece en el catálogo de la editorial.

8. Usa el formato sugerido en el apéndice A de selección de pruebas para hacer este ejercicio. Las posibles áreas en que puede hacerse incluyen pruebas de lectura diagnóstica para alumnos de escuelas primarias, pruebas de pensamiento creativo para alumnos universitarios, pruebas de admisión a universidades, medidas de depresión, ansiedad o trastornos alimentarios.

9. En distintos sitios de internet aparecen largas listas de pruebas psicológicas que incluyen pruebas de inteligencia y variados rasgos de personalidad. ¡Ten cuidado! Muchas de ellas no cuentan con información sobre confiabilidad, validez o normas: justo las características técnicas que abordamos en los tres capítulos siguientes. **No** recomendamos a los estudiantes consultar este tipo de listas; sin embargo, el estudiante puede querer entrar a alguno de estos sitios para ver "qué hay ahí afuera". Ejemplos de estos sitios son:

- AllTheTests, http://www.allthetests.com/
- PsychTests Test Yourself, http://testyourself.psychtests.com/
- Queendom, http://www.queendom.com/
- Quincy's Online Psychological and Personality Tests, http://www.quincyweb.net/quincy/psychology.html

El siguiente sitio tiene un conjunto de pruebas, empleadas principalmente con propósitos de investigación, que se pueden inspeccionar:

- http://www.yorku.ca/rokada/psyctest/

CAPÍTULO 3
Normas

Objetivos

1. Definir el propósito de las normas de las pruebas.

2. Refrescar tu conocimiento sobre los siguientes temas de estadística descriptiva: tipos de escalas, distribuciones de frecuencia, términos para describir las formas de las distribuciones, medidas de tendencia central, medidas de variabilidad y puntuaciones z.

3. Identificar los distintos tipos de puntuaciones naturales de las pruebas.

4. Identificar el significado de theta.

5. Definir estos tipos de normas: percentiles y rangos percentiles, puntuaciones estándar, normas de desarrollo.

6. Resumir las fortalezas y debilidades de cada tipo de norma.

7. Identificar las características de cada uno de estos tipos de grupos normativos: nacional, por conveniencia, del usuario, de subgrupo, local e institucional.

8. Distinguir entre la interpretación orientada al criterio e interpretación orientada a la norma.

9. Describir cómo determinar la utilidad de un grupo normativo.

Objetivo de las normas

Matt tuvo 36 respuestas correctas en una prueba de vocabulario; ¿eso es bueno o malo? Meg respondió "Sí" a 14 de 30 reactivos en una prueba de ansiedad; ¿eso significa que es excepcionalmente ansiosa, o despreocupada, para contestar una prueba? Dan tuvo 52 respuestas correctas en una prueba de lectura de 80 reactivos y 24 correctas en una prueba de ciencia de 40 reactivos; ¿es relativamente mejor en lectura o en ciencia? Este tipo de preguntas son las que se abordan cuando se habla de **normas de una prueba**. La idea básica es traducir lo que se denomina puntuación natural a otro tipo de puntuación normativa o estandarizada. La **puntuación natural** es el resultado más o menos inmediato de las respuestas de un individuo en una prueba. Para obtener la **puntuación normativa**, la puntuación natural de un individuo se *compara con las puntuaciones de individuos del grupo del que se obtuvieron las normas*, es decir, el grupo normativo o de estandarización. También se conoce a las puntuaciones normativas como puntuaciones *derivadas* o *escalares*. Este capítulo se ocupa de las dos preguntas cruciales en relación con las normas. Primera, ¿cuáles son los tipos de normas que se usan por lo general? Segunda, ¿qué son los grupos de estandarización, es decir, de dónde vienen las normas?

> Las puntuaciones naturales y las puntuaciones de la TRR (teoría de la respuesta al reactivo) son difíciles de interpretar sin más información, pero esto puede ser más fácil si dichas puntuaciones se convierten en *puntuaciones escalares*.
>
> *Standards*... (AERA, APA, & NCMW, 2013)[1]

Las nociones fundamentales de las normas no son exclusivas del campo de las pruebas psicológicas. Usamos las nociones de puntuación natural y posición relativa dentro de un grupo en numerosas situaciones cotidianas. Consideremos estos ejemplos: ¿1.95 m es alto? No lo es si se trata de un árbol, pero, si se trata de un ser humano, sí lo es. Incluso entre seres humanos, no es una estatura impresionante para un jugador de basquetbol de la NBA, pero es asombrosa para un alumno de sexto grado. ¿Y qué hay de un pulso de 100? No es extraordinario para un recién nacido (humano) ni para un adulto que acaba de hacer un ejercicio intenso; sin embargo, podría ser una señal de peligro para un adulto que se encuentra en reposo, ¿Qué piensas de correr 100 metros en 58.85 segundos? Para un equipo de atletismo de bachillerato o universitario, éste sería un pésimo tiempo, pero Ida Keeling corrió en este tiempo en junio de 2012 e impuso récord en EUA en la categoría de 95 o más años de edad: ¡excelente, Ida! (masterstrack.com, 2012). Estos ejemplos ilustran la idea de hacer comparaciones dentro de un grupo para realizar muchos tipos de interpretaciones en la vida cotidiana. En evaluación psicológica, operacionalizamos estas comparaciones en forma de normas.

Aquí hay otros ejemplos de puntuaciones naturales que constituyen el contexto de la

discusión de normas que aparece más adelante en este capítulo.

- Matt tuvo *36* reactivos correctos en una prueba de memoria a corto plazo. ¿Qué tan buena es su memoria?
- Meg respondió "Sí" a *14* reactivos de una medida de ansiedad. ¿Es despreocupada o un caso perdido para contestar pruebas?
- Dan tuvo *32* reactivos correctos en una prueba de lectura y *24* correctos en una prueba de matemáticas. ¿Es mejor en lectura o en matemáticas?
- Sheri mide *1.57* m. ¿Su estatura es alta, baja o promedio?
- El pulso de Tom es de *71* latidos por minuto. ¿Es normal?
- La pequeña Abigail fue calificada con *7* en una escala de afabilidad de 10 puntos. ¿Es un encanto o una niña antipática?

Repaso de estadística: parte 1

Las normas de una prueba se basan en nociones elementales de estadística descriptiva. Suponemos que el lector ha tomado algún curso introductorio de estadística, pero que puede necesitar refrescar en su mente algunos temas de ese curso. Esta sección ofrece un repaso; no pretende repetir el curso entero de estadística, ni enseñar cómo hacer cálculos estadísticos, sino ser un repaso rápido. Esta sección trata ideas clave de estadística univariada. El capítulo 4 incluye un repaso de estadística bivariada, es decir, correlación y regresión.

Los ejemplos de esta sección recurren a pequeños conjuntos de datos para ilustrar los conceptos estadísticos. En la práctica, por lo general, tenemos conjuntos mucho más grandes, a menudo de miles de casos, por lo que usamos programas de cómputo para aplicar la estadística a estos datos. En el apéndice D se presentan varios conjuntos de datos que el lector puede usar para calcular cualquiera de los estadísticos que revisamos en esta sección.

Variables

¿Cómo se relaciona la estadística con el campo de las pruebas psicológicas? Para contestar esta pregunta, tenemos que comprender los matices del significado del término variable. Una ciencia se construye alrededor de las variables que estudia; por ejemplo, la psicología estudia variables como inteligencia, extroversión, inadaptación o agudeza visual. Los objetos de estudio (humanos, ratas, etc.) varían a lo largo de estas variables de bajo a alto, de menos a más, o de maneras semejantes. Las variables se pueden describir en tres niveles de generalidad. En el nivel más general, una variable es un constructo, que consiste en descripciones verbales y definiciones de la variable; por ejemplo, la inteligencia puede definirse como la capacidad para manipular símbolos abstractos, la inadaptación puede describirse como un sentimiento o como evidencia objetiva de dificultades importantes para llevar a cabo actividades de la vida cotidiana.

En el segundo nivel, **medimos** la variable, lo cual constituye una **definición operacional.** El campo de las pruebas psicológicas se ocupa de estas medidas; estudia sus características y cataloga las medidas existentes. En el tercer nivel, obtenemos *datos crudos*, que son lo números resultantes de la aplicación de la medida.

Las estadísticas trabajan con datos crudos, el nivel más específico de una variable; ya que estos datos provienen de nuestras medidas, la estadística ofrece resúmenes y procesamientos de las medidas (pruebas) que nos interesan. Desde luego, en todo momento estamos más interesados en la variable al nivel de constructo.

La estadística se divide principalmente en descriptiva e inferencial; en la mayoría de las investigaciones, obtenemos una gran cantidad de datos crudos. Por ejemplo, podemos tener muchas diferentes puntuaciones obtenidas en una prueba por de cientos de individuos. La **estadística descriptiva** ayuda a resumir o describir estos datos con el fin

de comprenderlos. Es más frecuente que los datos provengan de una muestra de individuos, pero estamos interesados en conocer la población a la que pertenece esa muestra. La **estadística inferencial** nos ayuda a sacar conclusiones, inferencias, sobre lo que probablemente es verdad acerca de la población con base en lo que descubrimos de la muestra.

Resumen de puntos clave 3-1

Tres niveles para definir una variable

Constructo	Definición general de la variable
Medida	Definición operacional, a menudo una prueba
Datos crudos	Números que resultan de la medida

Tipos de escalas

Las variables se miden en escalas. Stevens (1951) presentó una clasificación en cuatro tipos de escalas que se cita con gran frecuencia en estadística y en el campo de las pruebas. Las distinciones entre estos tipos de escalas son importantes porque podemos realizar ciertos tipos de operaciones aritméticas y estadísticas con algunas, aunque con otras no.

La clasificación de las escalas se ciñe a su nivel de sofisticación. El nivel menos sofisticado o más primitivo es la **escala nominal**, que distingue simplemente objetos etiquetando cada uno de ellos con un número. Los números no significan más o menos, más grande o más pequeño, ni cualquier otra distinción cuantitativa. Ejemplos de esto son los números en las playeras de los jugadores de basquetbol, el número de seguridad social o los códigos 0 y 1 que se asignan a hombres y mujeres con el objetivo de codificarlos en una computadora.

En una **escala ordinal** se asignan números a los objetos para indicar un orden, como tener más o menos de un rasgo, sin especificar de ninguna manera las distancias entre dichos objetos en la escala. El ordenamiento por rangos ilustra las escalas ordinales; por ejemplo, las encuestas de preferencia de equipos de fútbol universitario ofrecen una clasificación de los equipos: 1, 2, 3, 4,... 25. Ya que se entiende que estos números están en una escala ordinal, no se puede inferir que la diferencia entre el equipo 2 y el 3 sea igual a la que hay entre el equipo 1 y el 2. De hecho, la diferencia entre los equipos 1 y 2 puede ser muy pequeña, mientras que entre los equipos 2 y 3 la diferencia puede ser muy grande.

La **escala intervalar** ubica los objetos en orden de menor a mayor y los intervalos entre distintos puntos de la escala son equivalentes. De ahí que, en una escala intervalar, la distancia entre 2 y 4 es la misma que entre 6 y 8 o entre 20 y 22. Sin embargo, esta escala carece de un **cero absoluto**; aunque tiene un cero, éste no indica la ausencia total de la variable que se mide. El ejemplo cásico de una escala intervalar es el termómetro que marca grados Fahrenheit; en esta escala, el cero *no* indica la ausencia total de calor.

Suma
y resta son operaciones legítimas en esta escala; por ejemplo, la diferencia entre 30° y 40° es la misma que la que existe entre 50° y 60°. Sin embargo, multiplicación y división no son operaciones legítimas, porque 60°F no representan el doble de calor que 30°F o la mitad de 120°F. Para hacer tales afirmaciones, sería necesario usar la escala Kelvin, que sí tiene un cero absoluto, el punto en que no existe el calor.

La **escala de razón**, la más sofisticada, ubica los objetos en orden de menor a mayor, con intervalos equivalentes, *y* tiene un cero absoluto. La mayoría de nuestras mediciones físicas, como altura y peso, son escalas de razón. En contraste, la mayoría de nuestras mediciones psicológicas son escalas ordinales o intervalares.

¿Cuál es la pertinencia de estas distinciones, en apariencia arcanas, para nuestro interés en la medición psicológica? Consideremos el caso de la conocida escala de CI. No es una escala de razón, pues no tiene un cero absoluto; de ahí que no sea legítimo decir que una persona con un CI de 150 es dos veces más inteligente que una persona con un CI de 75. Además, si la escala del CI es sólo ordinal, tampoco podemos decir que la diferencia entre un CI de 80 y otro de 100 es igual que la diferencia entre un CI de 120 y otro de 140. Tenemos que preocuparnos por la naturaleza de nuestras escalas para hacer declaraciones coherentes sobre nuestras mediciones o, en otras palabras, evitar decir estupideces sobre ellas. El método de Stevens (1951) para clasificar escalas, a pesar de ser tan citado, no es el único modo de pensar acerca de ellas, pero es muy útil.

Resumen de puntos clave 3-2

Resumen de los tipos de escalas

Escala	Propiedades
Nominal	Clasifica, asigna números
Ordinal	+ Ubica de acuerdo con la magnitud
Intervalar	+ Intervalos iguales en la escala
De razón	+ Hay un cero verdadero

Organización de datos crudos

Cuando nos encontramos con una gran cantidad de datos, a menudo queremos organizarlos. La manera más usual de hacerlo es mediante una **distribución de frecuencia**, la cual organiza los datos en grupos de puntuaciones cercanas. La figura 3-1 presenta un conjunto de datos y su distribución de frecuencia. Es difícil darle sentido a una serie de datos crudos, pero la distribución de frecuencia revela con facilidad características como el rango de puntuaciones y el área donde se concentran. La

distribución de frecuencia de la figura 3-1 muestra los intervalos en que se clasifican las puntuaciones, la frecuencia o conteo de las puntuaciones que corresponden a cada intervalo y la frecuencia acumulativa (f acu), es decir, la frecuencia en cada intervalo *y* debajo de éste.

Puntuaciones (N = 100)									
102	110	130	99	127	107	113	76	100	89
120	92	118	109	135	116	103	150	91	126
73	128	115	105	112	138	100	158	103	86
114	134	117	91	92	98	96	125	155	88
129	108	99	83	149	103	95	95	82	133
91	121	103	147	110	90	122	94	124	65
71	101	93	88	78	105	145	90	94	87
102	79	99	111	117	98	115	112	116	80
96	114	106	111	119	101	123	109	132	100
105	77	100	131	106	108	113	143	102	88

Distribución de frecuencia

Intervalo	F	F Acu
150-159	3	100
140-149	4	97
130-139	7	93
120-129	10	86
110-119	18	76
100-109	23	58
90-99	19	35
80-89	9	16
70-79	6	7
60-69	1	1

Figura 3-1. Muestra de un conjunto de puntuaciones naturales y una distribución de frecuencia.

Una distribución de frecuencia a menudo se convierte en una gráfica; las dos formas gráficas más comunes son el **histograma de frecuencias** y el **polígono de frecuencias**. La figura 3-2 presenta estas gráficas empleando los datos de la distribución de frecuencia de la figura 3-1.

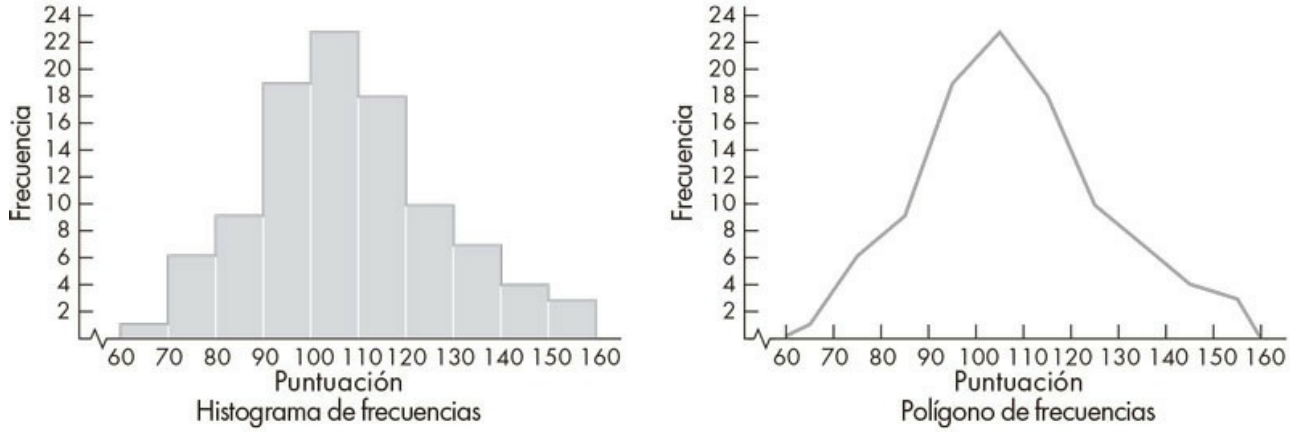


Figura 3-2. Histograma de frecuencias y polígono de frecuencias.

Tendencia central

Aunque la distribución, el histograma y el polígono de frecuencias son resúmenes muy útiles de los datos crudos, es conveniente tener un índice que represente mejor el conjunto entero de datos. Tal índice se denomina medida de **tendencia central**: el centro alrededor del cual los datos crudos tienden a agruparse. Existen tres medidas de tendencia central que se usan de manera habitual: media, mediana y moda.

La **media** es un promedio aritmético. Se representa con *M* o *X* (léase "equis-barra" o simplemente "media"). Su fórmula es:

$$M = \frac{\Sigma X}{N}$$

Fórmula 3-1

donde

X = puntuación o dato crudo

N = número de puntuaciones y

Σ = es el signo de suma, pues indica "suma todas estas cosas (representadas por *X*)".

La **mediana** es la puntuación que se encuentra justo a la mitad cuando las puntuaciones están ordenadas de menor a mayor. La mediana divide el conjunto de puntuaciones justo a la mitad. La **moda** es la puntuación que se presenta con mayor frecuencia. La figura 3-3 muestra ejemplos de las tres medidas de tendencia central de un pequeño conjunto de datos.

Datos crudos: 2, 5, 3, 6, 6, 7, 6, 4, 8, 3, 4
Media: $M = \Sigma X / N = 54 / 11 = 4.91$
Mediana: 5
Moda: 6

Figura 3-3. Medidas de tendencia central de un pequeño conjunto de datos.

¡Inténtalo!

Calcula las tres medidas de tendencia central de este conjunto de puntuaciones.
Puntuación:

2 5 4 5 3 4 5 2 6
Media = _____ Mediana = _____ Moda = _____

Variabilidad

Las medidas de tendencia central ofrecen un resumen muy útil de los datos, pero no nos dicen nada acerca de de su **variabilidad**. Por ejemplo, dos conjuntos de datos pueden tener la misma media, pero en uno todas las puntuaciones están a dos puntos de la media, mientras que en otro las puntuaciones pueden estar muy dispersas. Por lo tanto, para describir mejor los datos crudos, ofrecemos un índice de variabilidad.

El índice de variabilidad más simple es el **rango**, que consiste en la distancia entre la puntuación más baja y la más alta. Para indicar el rango, se puede especificar la puntuación más baja y la más alta o sólo la diferencia entre ellas. En el caso de los datos de la figura 3-1, podemos decir que el rango es 65-158 o 93.

La **desviación estándar** es el índice de variabilidad más usado. Se puede representar en distintos contextos con cualquiera de estos símbolos[2]: *D*, *DE* o σ. Su fórmula es:

$$DE = \sqrt{\frac{\sum (X - M)^2}{N}}$$

Fórmula 3-2

A menudo se escribe también como:

$$DE = \sqrt{\frac{\sum x^2}{N}}$$

Fórmula 3-3

donde $x = X - M$. Estas fórmulas son la definición formal de la desviación estándar. Cuando se calcula la *DE* de una muestra con el fin de tener una estimación de la *DE* de la población, la "*N*" de la fórmula se reemplaza con "$N - 1$".

La fórmula no ayuda a formarnos una idea intuitiva de cómo es que la *DE* mide la variabilidad. Sin embargo, tiene felices consecuencias en términos matemáticos para su uso posterior, así que tan sólo hay que acostumbrarse a su presencia desgarbada. La desviación estándar se usará en gran medida más adelante en este capítulo.

Un índice de variabilidad que tiene una relación muy estrecha es la **varianza**, la cual es simplemente la *DE* elevada al cuadrado; a la inversa, la *DE* es la raíz cuadrada de la varianza. En algunos libros avanzados de estadística, la varianza se usa más que la desviación estándar, pero en el campo de las pruebas psicológicas, usamos la desviación

estándar con mayor frecuencia, aunque no siempre. La figura 3-4 muestra la desviación estándar (*DE*), la varianza y el rango de un pequeño conjunto de datos.

Puntuaciones:	6	9	4	5	5	1	$M = X/N = 30/6 = 5$
$(X - M) = x$	1	4	-1	0	0	-4	
	$x^2 = 1$	16	1	0	0	16	
$DE = 2.38$	Varianza = $DE^2 = 5.66$						Rango = 1 – 9

Figura 3-4. Medidas de variabilidad de un pequeño conjunto de datos.

Una cuarta medida de variabilidad es el **rango intercuartílico**. Como lo sugiere su nombre, es la distancia entre el primero y el tercer cuartil. El primero y el tercer cuartil corresponden a los percentiles 25 y 75, respectivamente. Si ya se conocen los percentiles, como suele ocurrir con los datos de las pruebas psicológicas, se puede determinar con facilidad el rango intercuartilar.

¡Inténtalo!

Calcula la desviación estándar y el rango de los siguientes datos.
Puntuaciones: 1 4 7
DE = _____ Rango = _____

Puntuaciones z

Una **puntuación z**, o simplemente *z*, se define como:

$$z = \frac{X - M}{DE}$$

Fórmula 3-4

donde *X* es una puntuación individual o un dato, *M* es la media y *DE* es la desviación estándar. La distribución de las puntuaciones z tiene una media de 0 y una *DE* de 1. De ahí que, sin importar los valores de las puntuaciones originales, cuando se convierten en puntuaciones z, siempre tienen la misma media y desviación estándar.

Las puntuaciones z se usan para "trazar" la curva normal en términos de áreas bajo la curva. La figura 3-5 ilustra este uso de las puntuaciones z. Recordemos las tablas de las áreas bajo la curva de la estadística básica. Debido a que las puntuaciones z tienen una

media (0) y una desviación estándar (1) en común, sin importar los valores de las puntuaciones originales, su papel es crucial en el desarrollo de las normas de las pruebas.

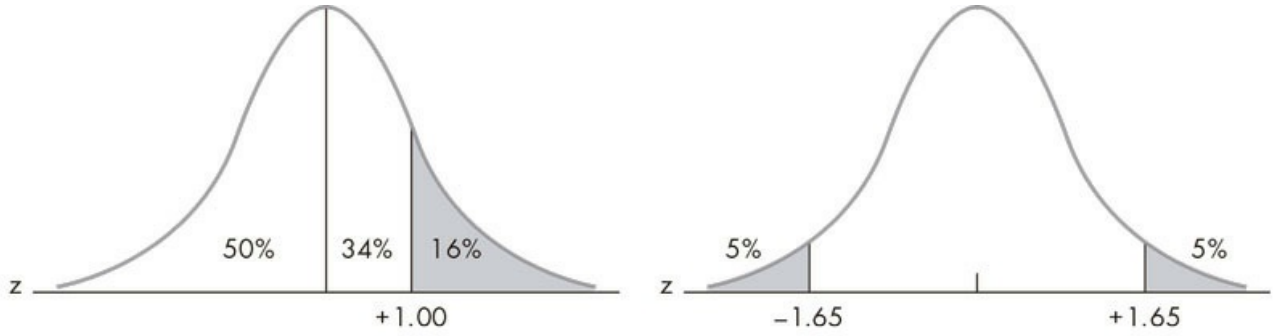


Figura 3-5. Ejemplos de las puntuaciones z que marcan áreas bajo la curva normal.

Formas de distribuciones

En el campo de las pruebas psicológicas, solemos hacer referencia a la forma de una distribución de frecuencias de las puntuaciones de una prueba, por lo que debemos conocer los términos empleados para describir estas formas. El punto de referencia o la distribución por excelencia es la **curva normal** o distribución normal. Su popular nombre es curva de campana, aunque el parecido con la campana no sea tan notable. La curva, generada por una fórmula difícil de manejar[3], es una función de densidad; es decir, el área bajo la curva está llena, en realidad repleta, de puntos que representan datos. La distribución de la figura 3-5 es una curva normal; está trazada con una media de 0 al centro y una escala establecida por las unidades de desviación estándar (a veces llamadas unidades σ) en su base. Observa con atención la posición de las unidades de desviación estándar.

Esta distribución es *unimodal*, pues sólo tiene una "joroba". Es *simétrica* en relación con su eje central, una línea imaginaria perpendicular a la base a la altura de la media. Alrededor de su eje central, el lado izquierdo de la curva es la imagen en espejo del derecho. Las "colas" de la curva son *asintótica*s en relación con la base, es decir, continúan hasta el infinito acercándose siempre a la base, pero sin llegar nunca a ella. Desde luego, esto es verdad sólo para el modelo teórico de la curva normal, pues en la práctica los datos terminan en algún punto finito. Nota que casi toda el área (cerca de 99.8%) bajo la curva está contenida dentro de +/− 3 unidades de σ.

Las distribuciones pueden "alejarse de la normalidad", es decir, ser diferentes de la curva normal de varias maneras. La figura 3-6 muestra estas desviaciones; la primera lo es en términos de la **curtosis**, que se refiere a qué tan pronunciado es el pico de la distribución. En una distribución leptocúrtica es más pronunciado el pico, mientras que una distribución platicúrtica es más plana que la distribución normal.

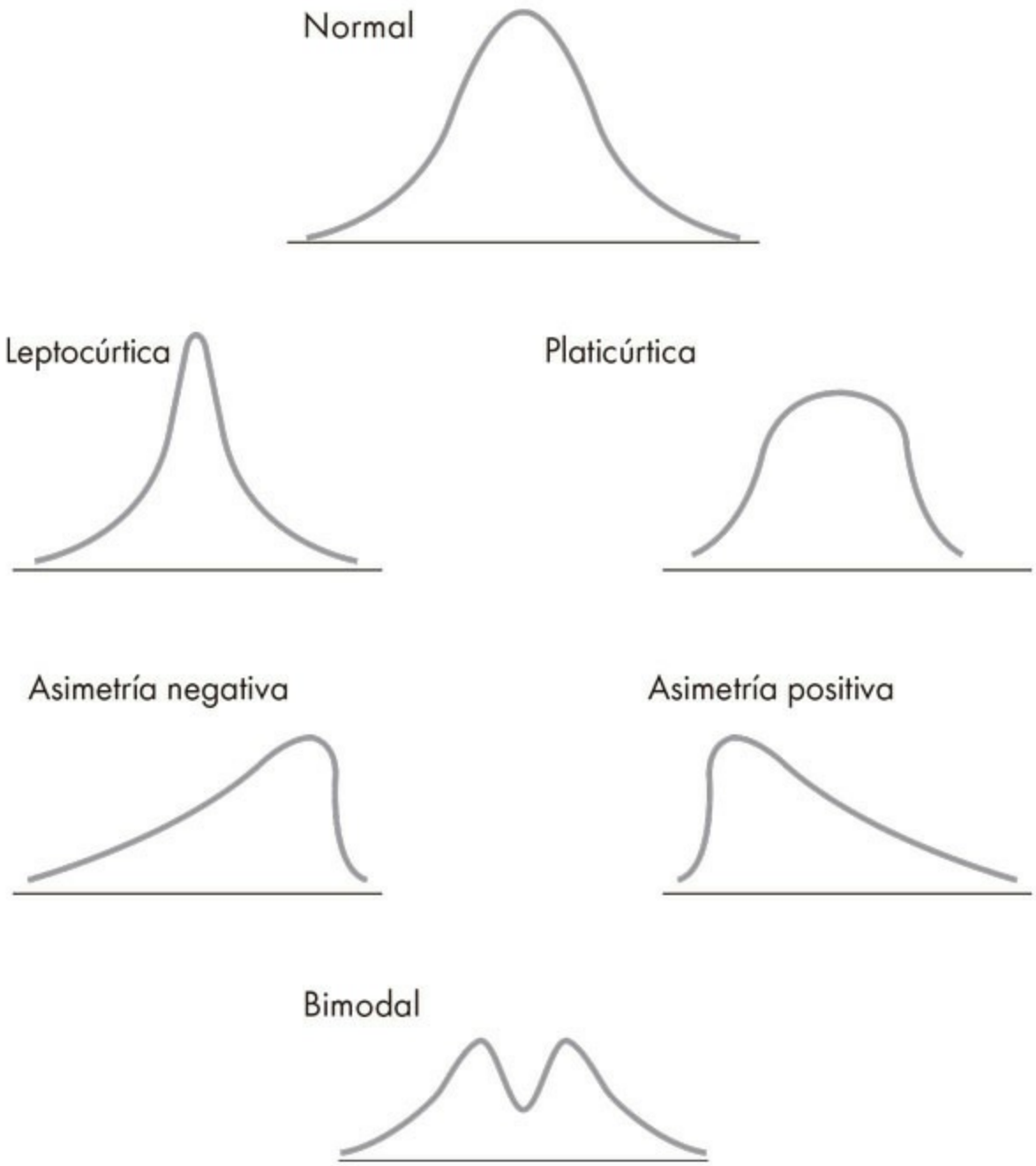


Figura 3-6. Ejemplos de varias distribuciones.

Otro tipo de desviación de la normalidad es en términos de **asimetría**, el grado en que los lados izquierdo y derecho de la curva son diferentes. La asimetría **negativa** o a la izquierda tiene una larga cola del lado izquierdo y un abultamiento de puntuaciones del lado derecho. La asimetría **positiva** o a la derecha tiene una larga cola del lado derecho y un abultamiento de puntuaciones del lado izquierdo. El último tipo de desviación de la normalidad es en términos de la modalidad de la distribución. Cuando es normal, la distribución es unimodal, pero puede tener más de una moda, como la distribución

bimodal que aparece en la parte inferior de la figura 3-6.

¿Por qué es importante la curva normal para las pruebas psicológicas? Muchos fenómenos que ocurren de manera natural tienden a tener una distribución normal; por ejemplo, en el caso de la gente de cierto sexo y edad, las siguientes variables, que pueden medirse con facilidad, tienen una distribución normal: estatura, memoria a corto plazo para los dígitos, tiempo de reacción, fuerza de agarre y así sucesivamente. Esta tendencia hacia la normalidad es tan omnipresente que, cuando la distribución real de una variable es desconocida, no es impensable suponer que probablemente sea normal. De ahí que muchas pruebas se construyan para obtener una distribución más o menos normal de las puntuaciones. Sin embargo, a veces creamos pruebas de manera deliberada para obtener distribuciones asimétricas en sentido positivo o negativo, como lo discutiremos en el capítulo 6.

También hay algunas distribuciones, que ocurren de manera natural, que son marcadamente no normales, por ejemplo, las distribuciones de ingresos, peso y población de la ciudad.

Con esta revisión de la estadística es suficiente para salir bien librados de este capítulo. Extenderemos esta revisión al principio del capítulo 4.

Puntuación natural

Todas las normas de las pruebas son transformaciones de las *puntuaciones naturales*; de ahí que antes de definir los tipos de normas, será de utilidad considerar estas puntuaciones. La **puntuación natural** es el resultado más inmediato que se obtiene de calificar una prueba, y puede aparecer de distintas maneras. Puede ser el número de respuestas correctas en una prueba de aprovechamiento, el número de preguntas respondidas en cierta dirección –por ejemplo, "Sí" o "De acuerdo" en un inventario de personalidad o intereses– o la suma de respuestas codificadas numéricamente en una serie de reactivos de actitud, como en una escala de actitud de 10 reactivos. En ésta, cada reactivo exige una respuesta en una escala de cinco puntos, que va desde En total desacuerdo (1) a En total acuerdo (5). La puntuación natural es la suma de respuestas numéricas a los 10 reactivos. La figura 3-7 muestra un ejemplo de este tipo de puntuación natural de una medida de seis reactivos de actitudes hacia las matemáticas. En este ejemplo, los reactivos con una redacción o connotación negativa invierten los valores de la escala con el fin de determinar la puntuación natural.

	Respuesta					
	En total desacuerdo				**En total acuerdo**	**(Puntuación del reactivo)**
Reactivo	1	2	3	4	5	
1. El álgebra es muy divertida.	[ ]	[X]	[ ]	[ ]	[ ]	(2)
2. La geometría es para los raros.	[ ]	[ ]	[ ]	[ ]	[X]	(1)
3. Me gusta hacer cálculos.	[X]	[ ]	[ ]	[ ]	[ ]	(1)
4. Las matemáticas son muy útiles para mí.	[ ]	[ ]	[X]	[ ]	[ ]	(3)
5. Me encanta la estadística.	[ ]	[X]	[ ]	[ ]	[ ]	(2)
6. Las ecuaciones me provocan escalofríos.	[ ]	[ ]	[ ]	[X]	[ ]	(2)
				Puntuación natural = 11		

Figura 3-7. Obtención de la puntuación natural de una escala de actitudes hacia las matemáticas.

Las medidas antropométricas y fisiológicas también pueden considerarse puntuaciones naturales. Sheri mide 1.57 m de estatura; el pulso de Dan es de 54 latidos por minuto; Grazia nada 180 m de mariposa en 2:20. Todas estas medidas son puntuaciones naturales; reemplazarlas en un contexto normativo ayudará a interpretarlas. Las normas ayudan a responder a preguntas como éstas: ¿Sheri es muy alta para su edad? ¿El pulso de Dan es anormal? ¿Grazia nada a nivel de competidora olímpica?

> Por lo general, la calificación empieza seleccionando las respuestas a los diferentes tipos de reactivos de la prueba. Las puntuaciones de los reactivos se combinan, a veces sumándolas, para obtener una *puntuación natural* utilizando la teoría clásica de las pruebas o para producir una puntuación TRR utilizando la teoría de la respuesta al reactivo...
>
> *Standards*... (AERA, APA, & NCME, 2013)

Los procedimientos de calificación de algunas pruebas requieren una puntuación natural "corregida" o "ajustada"«49a. El más popular de estos ajustes es la *corrección por adivinación*. Ésta se aplica en algunas pruebas de capacidad y aprovechamiento que emplean un formato de respuesta de opción múltiple. La teoría dice que en este formato se pueden obtener respuestas correctas adivinando a ciegas; es decir, se puede obtener 1 / K reactivos correctos sólo por adivinar la respuesta, donde K es el número de opciones de respuesta posibles. Así, la fracción $[I/(K - 1)]$, donde I es el número de respuestas incorrectas, se resta de la puntuación natural original para obtener una puntuación natural corregida. La corrección por adivinación, que alguna vez fue muy empleada, ha caído en desuso.

El caso especial de theta (θ) «49-50

Recordemos la discusión de la teoría de la respuesta al reactivo (TRR) del capítulo 1. Mientras que una prueba calificada de la manera convencional proporciona una puntuación natural que es la suma de las respuestas a todos los reactivos de la prueba, una prueba calificada de acuerdo con la TRR no es una simple suma de respuestas, sino una función de las *respuestas del examinado que interactúan con las características de los reactivos*. La puntuación TRR suele llamarse **theta** (θ). Consideremos dos ejemplos de cómo θ puede surgir; en ambos, es *crucial* que los reactivos sean seleccionados de acuerdo con la unidimensionalidad y escalados según su dificultad en programas de investigación previos. Trataremos las características TRR de los reactivos de manera más completa en el capítulo 6. La calificación real de las pruebas de acuerdo con los procedimientos TRR es compleja, pues requiere de sofisticados programas de cómputo. No es una simple cuestión de sumar las respuestas correctas. Sin embargo, en este punto, podemos dar una idea aproximada de cómo la metodología TRR proporciona puntuaciones theta.

Primero, consideremos los datos de la figura 3-8; los reactivos están ordenados de acuerdo con su dificultad de izquierda a derecha. Los reactivos con números inferiores son muy fáciles, mientras que los que tienen números superiores son muy difíciles. Por ejemplo, si ésta fuera una prueba de cálculo aritmético para estudiantes de escuelas primarias, el reactivo 1 podría ser 6 + 3; el reactivo 10, 740 – 698; y el reactivo 20, 0.56 × 1.05. Aquí hemos clasificado los reactivos simplemente como fáciles, moderados y difíciles. En una aplicación real se usarían los valores de dificultad exactos. (Los reactivos no necesitan ordenarse de acuerdo con la dificultad físicamente, pero es útil mostrar tal ordenamiento en nuestros ejemplos.) Cada "x" en la figura representa una

respuesta correcta. A Miguel se le presentaron los reactivos 1 al 10 y obtuvo 7 aciertos, mientras que a José se le aplicaron los reactivos 11 al 20 y también obtuvo 7 aciertos. Ya que José respondió a más reactivos difíciles, se le asigna una puntuación theta más alta a pesar de que ambos examinados tuvieron 7 aciertos. Obviamente, este procedimiento no funcionará a menos que los reactivos primero se ordenen de acuerdo con su dificultad. La figura 3-9 muestra cómo las respuestas de la figura 3-8 se traducen a puntuaciones de la escala theta alineada con los reactivos.

Reactivo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Dificultad del reactivo			Fácil							Moderada						Difícil				
Miguel	x	x	x	x		x		x		x										
José											x	x	x	x		x		x	x	

x = respuesta correcta

Figura 3-8. Derivación de una puntuación theta de diferentes conjuntos de reactivos.

Reactivo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Dificultad del reactivo			Fácil							Moderada						Difícil				
Valores theta	−4.0		−3.0		−2.0		−1.0			0.0			+1.0		+2.0		+3.0		+4.0	
Puntuación de Miguel					Θ= −2.0															
Puntuación de José															Θ = +2.0					

Figura 3-9. Puntuaciones theta alineadas con los reactivos ordenados de acuerdo con su dificultad.

¡Inténtalo!

En internet hay excelentes demostraciones de la interacción de la respuesta del examinado con las características del reactivo empleando la metodología TRR. Entra a http://edres.org/scripts/cat/catdemo.htm. Entra al sistema y trata de responder la prueba. Observa cómo la "estimación de la capacidad" se ajusta después de cada respuesta. Aunque las puntuaciones se reportarán como percentiles o alguna otra puntuación escalar o normativa, pueden transformarse de theta, que se actualiza después de cada respuesta.

Theta tiene algunas propiedades de la puntuación natural y otras de la puntuación escalar, es como una puntuación natural porque también es un resultado relativamente inmediato de las respuestas del examinado. Además, como una puntuación natural, no tiene significado por sí misma. Es como una puntuación escalar porque no es una simple suma de las respuestas del examinado; no depende sólo de que la respuesta sea correcta

(o “Sí” o respuestas como ésa), sino también de los valores TRR del reactivo al que corresponde la respuesta. Theta ubica al examinado de acuerdo con un rasgo o capacidad que, se supone, subyace en el conjunto total de reactivos del banco de la prueba. No es fácil interpretar los valores numéricos de la dimensión, porque son arbitrarios; sin embargo, suelen establecerse con 0.00 como valor central y un rango aproximado de −4.0 a +4.0, por lo que pueden verse como puntuaciones z. No obstante, estos valores no se refieren a una posición dentro de un grupo de normalización bien definido.

Aunque los valores theta pueden ser el resultado inmediato de las respuestas del examinado en un contexto TRR, en la práctica suelen interpretarse por medio de las puntuaciones escalares o normativas que se presentan en este capítulo. Es decir, se convierten en percentiles, puntuaciones estándar, equivalentes de grado y así sucesivamente. De hecho, en un informe de puntuaciones de una prueba, podemos encontrar un percentil o una puntuación T sin darnos cuenta de que es una transformación de una puntuación theta en vez de que lo sea de una puntuación natural convencional.

Tipos de normas

En esta sección se discuten los tipos de normas que, por lo general, se usan con las pruebas psicológicas. Hay tres principales categorías de normas, con varias subcategorías; en esta sección, describiremos cada una de ellas y señalaremos sus fortalezas y debilidades. En el caso de muchas pruebas, se dispone de varios tipos de normas, pero casi siempre se relacionan entre sí de manera sistemática. De modo que podemos convertir un tipo de norma en otro, aunque esto no sea posible con todos los tipos de normas. Las relaciones entre las normas son importantes; por lo general, se conceptualizan en el contexto de la curva normal que revisamos antes en este capítulo. Varias de estas relaciones se muestran en las figuras 3-10a y 3-10b y se representan de modo tabular en el cuadro 3-1.

La figura 3-10a es clásica, pero un poco abrumadora, mientras que la figura 3-10b es una versión simplificada. El principiante o novato hará bien en aprenderse de memoria esta versión simplificada, pues los psicólogos experimentados, que enfrentan la necesidad de integrar de manera rápida información de varias fuentes, a menudo usan una imagen mental como la de la figura 3-10b para hacer comparaciones entre las puntuaciones de distintos sistemas de normas. Desde luego, el cuadro 3-1 permite hacer conversiones exactas cuando sea necesario.

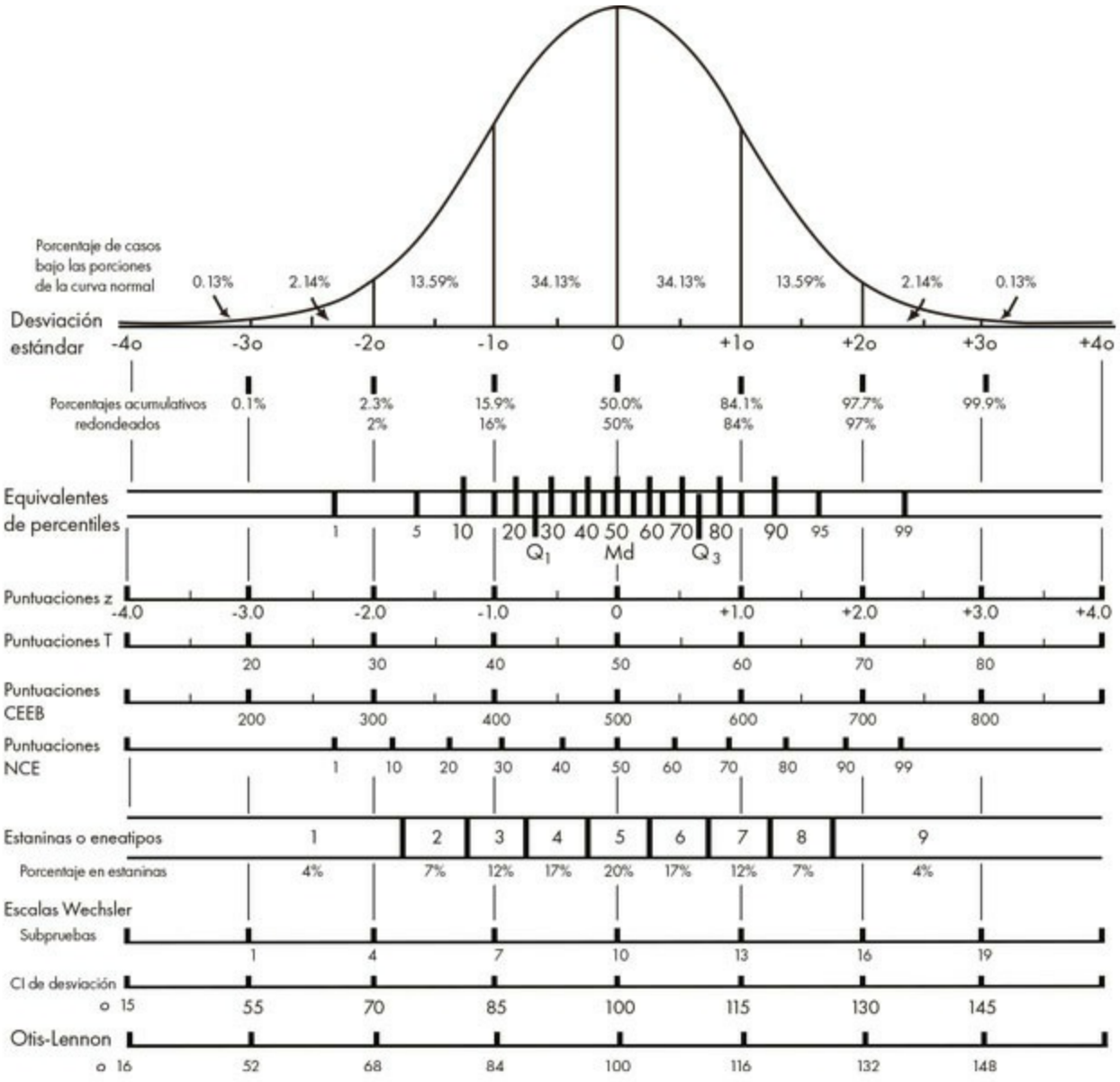
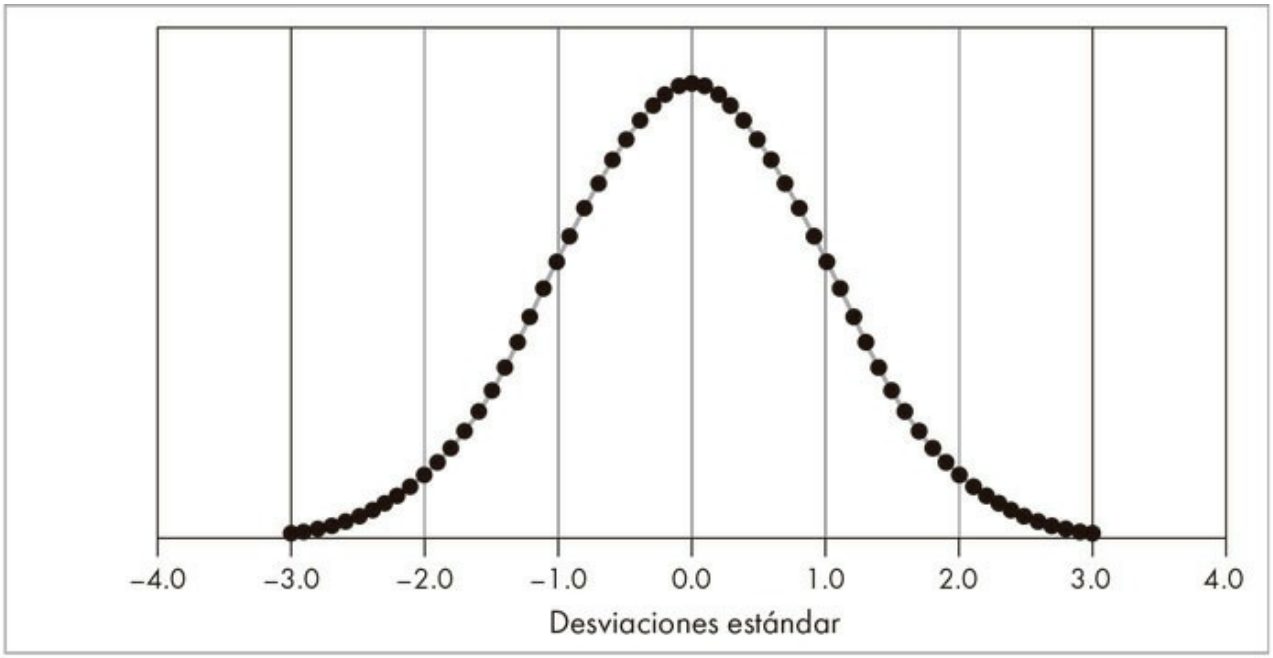


Figura 3-10a. Equivalencia de distintos tipos de normas en la curva normal.
Fuente: Tomado de Seashore, H. G. *Test service notebook 148: Methods of expressing test scores.*
Con autorización de la editorial Psychological Corporation.



Puntuaciones z	-3	-2	-1	0	+1	+2	+3
Percentiles	<1	2	16	50	84	98	>99
Puntuaciones T	20	30	40	50	60	70	80
CI de desviación	55	70	85	100	115	130	145
SAT	200	300	400	500	600	700	800

Figura 3-10b. Versión simplificada de las comparaciones entre sistemas de normas de una distribución normal.
Fuente: Adaptada de *Clinician's guide to evidence based practices: Mental health and addiction* por Norcross, Hoigan and Koocher (2008) fig. 5.1, pag. 126. Con autorización de Oxford University Press, USA

Cuadro 3-1. Equivalentes con percentiles de distintos sistemas de puntuaciones estándar

Percentil	Estanina o eneatipo	NCE	CI(15)	CI(16)	W-sub	Puntuación T	SAT	Puntuación z	Percentil
99	9	99	133	135	17	74	740	2.4	99
98	9	93	130	132	16	70	700	2.0	98
97	9	90	129	130		69	690	1.9	97
96	9	87	127	128		68	680	1.8	96
95	8	85	125	126	15	66	660	1.6	95

94	8	83	123	125					94
93	8	81	122	124		65	650	1.5	93
92	8	80	121	122		64	640	1.4	92
91	8	78	120	121					91
90	8	77	119		14	63	630	1.3	90
89	8	76		120					89
88	7	75	118	119		62	620	1.2	88
87	7	74	117	118					87
86	7	73	116	117		61	610	1.1	86
85	7	72							85
84	7	71	115	116	13	60	600	1.0	84
83	7	70							83
82	7	69	114			59	590	0.9	82
81	7	68	113	114					81
80	7	68							80
79	7	67	112	113		58	580	0.8	79
78	7	66							78
77	7	66	111	112					77
76	6	65				57	570	0.7	76
75	6	64	110	111	12				75
74	6	64							74
73	6	63	109	110		56	560	0.6	73
72	6	62							72
71	6	62		109					71
70	6	61	108						70
69	6	60		108		55	550	0.5	69
68	6	60	107						68
67	6	59		107					67
66	6	59	106			54	540	0.4	66
65	6	58		106					65
64	6	58							64
63	6	57	105		11				63
62	6	56		105		53	530	0.3	62
61	6	56	104						61
60	6	55		104					60
59	5	55							59
58	5	54	103	103		52	520	0.2	58

57	5	54							57
56	5	53							56
55	5	53	102	102					55
54	5	52				51	510	0.1	54
53	5	52	101						53
52	5	51		101					52
51	5	50							51
50	5	50	100	100	10	50	500	0.0	50
49	5	50							49
48	5	49		99					48
47	5	48	99						47
46	5	48				49	490	−0.1	46
46	5	48				49	490	−0.1	46
45	5	47	98	98					45
44	5	47							44
43	5	46							43
42	5	46	97	97		48	480	−0.2	42
41	5	45							41
40	5	45		96					40
39	4	44	96						39
38	4	44		95		47	470	−0.3	38
37	4	43	95		9				37
36	4	42							36
35	4	42		94					35
34	4	41	94			46	460	−0.4	34
33	4	41		93					33
32	4	40	93						32
31	4	40		92		45	450	−0.5	31
30	4	39	92						30
29	4	38		91					29
28	4	38							28
27	4	37	91	90		44	440	−0.6	27
26	4	36							26
25	4	36	90	89	8				25
24	4	35				43	430	−0.7	24
23	4	34	89	88					23
22	3	34							22

21	3	33	88	87		42	420	−0.8	21
20	3	32							20
19	3	32	87	86					19
18	3	31	86			41	410	−0.9	18
17	3	30		85					17
16	3	29	85	84	7	40	400	−1.0	16
15	3	28							15
14	3	27	84	83		39	390	−1.1	14
13	3	26	83	82					13
12	3	25	82	81		38	380	−1.2	12
11	3	24		80					11
10	2	23	81		6	37	370	−1.3	10
9	2	22	80	79					9
8	2	20	79	78		36	360	−1.4	8
7	2	19	78	76		35	650	−1.5	7
6	2	17	77	75					6
5	2	15	76	74	5	34	340	−1.6	5
4	2	13	74	72		32	320	−1.8	4
3	1	10	72	70		31	310	−1.9	3
2	1	7	70	68	4	30	300	−2.0	2
1	1	1	67	65	3	29	290	−2.4	1

CI(15) es para las pruebas de CI cuya M = 100 y DE = 15 (p. ej., la puntuación total de Wechsler o la de Stanford-Binet 5a. ed.). CI(16) es para las pruebas de CI cuya *M* = 100 y *DE* = 16, como SB 4a. ed. y Otis-Lennon. SAT cubre cualquier prueba cuya *M* = 500 y *DE* = 100. W Sub representa las subpruebas Wechsler y de SB5, cuya *M* = 10 y *DE* = 3

Fuentes: Psychologist Desk reference 2nd edition por Koocher, Norcross y Hill (2005). Adaptado del cuadro 1, pag. 111-116 © 1998 por Gerald P. Koocher, John C. Norcross y Sams, Hill, con autorización de Osford University Press, USA

Resumen de puntos clave 3-3

Principales categorías de las normas de las pruebas

- Rangos percentiles
- Puntuaciones estándar
- Normas de desarrollo

Rangos percentiles y percentiles

Uno de los tipos más comunes de normas en el caso de las pruebas psicológicas es el rango percentil o percentil. Existe una distinción técnica entre ambos términos. El **rango percentil** (RP) nos dice el porcentaje de casos en el grupo normativo que están por debajo de cierta puntuación natural. Así, si una puntuación natural de 48 tiene un RP de 60, significa que 60% de los casos de dicho grupo tuvo una puntuación natural igual o menor de 48. Esta puntuación se considera como un intervalo que va de 47.5 a 48.5, por lo que 48 está en la mitad del intervalo. Por ello, en algunas aplicaciones, el RP se calcula para que incluya la mitad de los casos que están en el intervalo de la puntuación natural.

Un **percentil** es un punto en una escala debajo del cual cae un porcentaje especificado de casos. La diferencia entre un percentil y un rango percentil puede resumirse del siguiente modo: en el caso del percentil, uno empieza con un porcentaje determinado y luego encuentra la puntuación natural que corresponde a este punto, mientras que en el caso de un rango percentil, uno empieza con una puntuación determinada y luego encuentra el porcentaje de casos que caen por debajo de esta puntuación. En la práctica, los términos *percentil* y *rango percentil* a menudo son intercambiables sin ningún problema.

A veces, encontramos ramificaciones del sistema de percentiles, entre las que se encuentran los deciles, quintiles y cuartiles. Como lo indican sus raíces latinas, estos sistemas dividen la distribución en diez, cinco y cuatro partes, respectivamente. En este sentido, podemos pensar en los percentiles como un sistema que divide en cien partes la distribución.

Las figuras 3-10a y 3-10b ilustran el lugar de los rangos percentiles en la curva normal, que varían desde 1 hasta 99 y 50 es su punto medio o mediana. La relación entre rangos percentiles y puntuaciones z está definida por el cuadro de áreas bajo la curva normal, el cual se estudia en estadística básica.

¡Inténtalo!

Usa la figura 3-10b para responder a estas preguntas. Escribe las estimaciones y la puntuación z que corresponda a cada rango percentil. Luego usa el cuadro 3-1 para revisar tus estimaciones.

RP	z estimada (Figura 3-10b)	z del cuadro (cuadro 3-1)
50	_____	_____
84	_____	_____
16	_____	_____
25	_____	_____
99	_____	_____

Fortalezas y debilidades de los rangos percentiles[4]

Los rangos percentiles tienen varias características atractivas; primero, es un concepto sencillo, fácil de entender y explicar, incluso a una persona sin conocimientos de estadística. También es fácil calcularlos a partir de un grupo de normativo. Por estas razones, los rangos percentiles son muy utilizados.

Los rangos percentiles tienen dos inconvenientes principales. Primero, las personas no expertas confunden con frecuencia el rango percentil con la **puntuación de porcentaje correcto** que se usa con muchas pruebas de salón de clases. De acuerdo con una tradición consagrada, en el sistema de puntuación de porcentaje correcto, 90% es A, 60% es fallido, y así sucesivamente. De este modo, un rango percentil de 72, que está por encima del desempeño promedio, puede ser malinterpretado como un desempeño apenas aprobatorio. Un rango percentil de 51, que es prácticamente el promedio, parece referirse a un desempeño deficiente en el sistema de porcentaje correcto. El psicólogo debe distinguir con cuidado entre el rango percentil y la puntuación de porcentaje correcto, sobre todo al interpretar las puntuaciones para personas no expertas.

La segunda desventaja de los rangos percentiles es la marcada desigualdad de las unidades en varios puntos de la escala. En particular, los rangos percentiles están "aglutinados" en la mitad de la distribución y "dispersos" en los extremos. Al principio, esta peculiaridad suena a tecnicismo trivial; sin embargo, tiene implicaciones prácticas importantes. Una diferencia en las puntuaciones naturales, digamos de 3 puntos, cubrirá muchos puntos percentiles en la mitad de la distribución, pero sólo unos pocos en los extremos. Este fenómeno puede observarse en las normas de rangos percentiles de cualquier prueba. Esta dificultad no es un atributo de los rangos percentiles, sino del hecho de que se apliquen a una variable que sigue una distribución normal, lo cual es difícilmente cierto en la mayoría de las pruebas psicológicas. La dificultad no se presentaría en circunstancias inusuales en las que la variable tenga una distribución rectangular; de hecho, el fenómeno se invertiría adoptando una distribución en forma de U.

La figura 3-11 ilustra el problema de unidades desiguales en el sistema de percentiles; se muestran las normas en percentiles de la escala de puntuaciones naturales. Puede observarse que un cambio en las unidades de puntuación natural de 7 a 10 (3 puntos) resulta en un cambio de sólo 2 puntos de percentil, mientras que un cambio en la puntuación natural de 17 a 20, también de 3 puntos, resulta en un enorme cambio de 24 puntos en unidades de percentil. Consideremos estos dos escenarios: Olivia gana 3 puntos en la puntuación natural de la primera a la segunda aplicación de una prueba, ¿qué tan grande es la diferencia en su posición en términos de percentiles? Eso depende de dónde se ubica en la distribución de puntuaciones. Si está en el extremo inferior o superior, sus 3 puntos más producen una pequeña diferencia en su posición, pero si se encuentra a la mitad de la distribución, sus 3 puntos más harán una gran diferencia en su posición. Un resultado similar ocurre si comparamos dos individuos cuyas puntuaciones naturales difieren por 3 puntos; si estos individuos se encuentran cerca de la parte inferior o superior de la distribución, sus posiciones en términos de percentiles serán similares, pero si se encuentran a la mitad de la distribución, sus posiciones serán notablemente

diferentes. Todos estos ejemplos se basan en una distribución más o menos normal, donde la "joroba" está en la mitad. En una distribución muy asimétrica, donde la "joroba" está a la izquierda o a la derecha del centro, los análisis que acabamos de describir cambian en consecuencia.

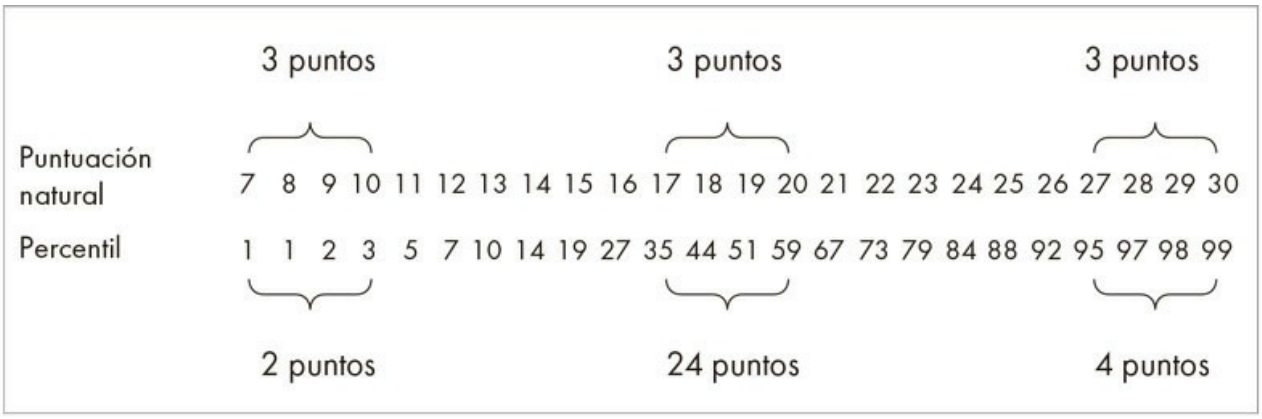


Figura 3-11. Normas de percentiles de una escala de puntuaciones naturales que muestran la desigualdad de las unidades percentiles.

Puntuaciones estándar

Las puntuaciones estándar son otro tipo de normas que se usan con frecuencia en las pruebas educativas y psicológicas. Conforman una familia de normas, pues hay varias versiones muy utilizadas de ellas y un número potencialmente infinito de otras versiones. Primero, describimos lo que tienen en común todas las puntuaciones estándar y, luego, identificamos las propiedades de las versiones específicas. Véanse los ejemplos del cuadro 3-2.

Cuadro 3-2. Algunos sistemas de puntuación estándar muy usados

Prueba	**Media**	**DE**
Escala Total de Wechsler y Stanford-Binet (SB)	100	15
Puntuaciones de subpruebas de Wechsler y SB	10	3
Law School Admissions Test	150	10
SAT	500	100
MMPI	50	10

Un sistema de **puntuación estándar** es una conversión de puntuaciones z (revisadas antes en este capítulo) en un nuevo sistema con una media (*M*) y una desviación estándar (*DE*) elegidas de manera arbitraria. Como *M* y *DE*, en este sistema suelen elegirse números lindos y fáciles de recordar como 50 y 10 o 500 y 100. En pocos casos, como veremos, se buscan otras características deseables.

Para convertir una puntuación natural en una estándar, primero se traduce la puntuación natural a una puntuación z; ésta se multiplica, después, por la nueva *DE* de la puntuación estándar y se suma la nueva media también de la puntuación estándar. En la figura 3-12 se esquematizan estos pasos. La siguiente fórmula realiza estos pasos:

$$PE = \frac{DE_e}{DE_n}(X - M_n) + M_e$$

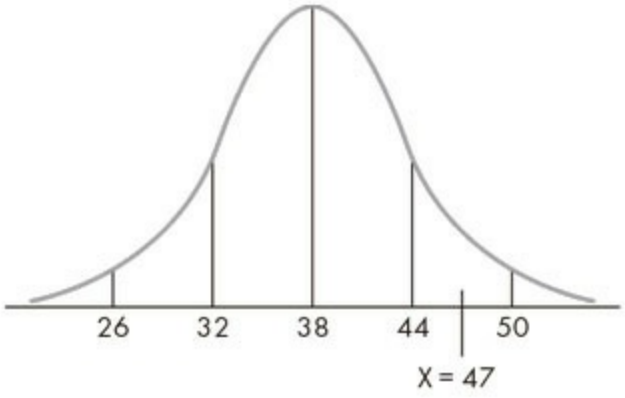
Fórmula 3-5

PE = puntuación estándar deseada
DE_e = desviación estándar en el sistema de puntuación estándar
DE_n = desviación estándar en el sistema de puntuación natural
M_n = media en el sistema de puntuación natural
M_e = media en el sistema de puntuación estándar
X = puntuación natural
Cuando la puntuación (X) se traduce a puntuación z, la fórmula es:

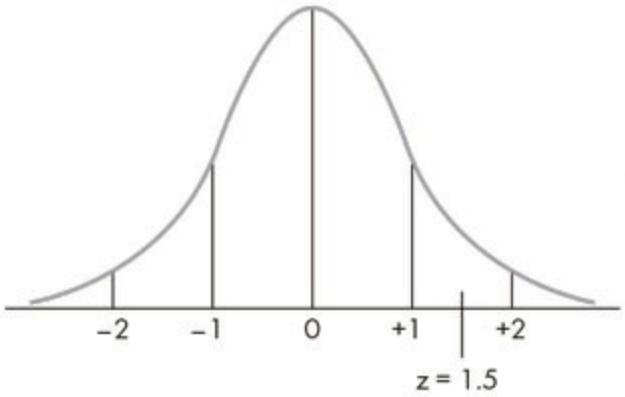
$$PE = z(DE_e) + M_e$$

Fórmula 3-6

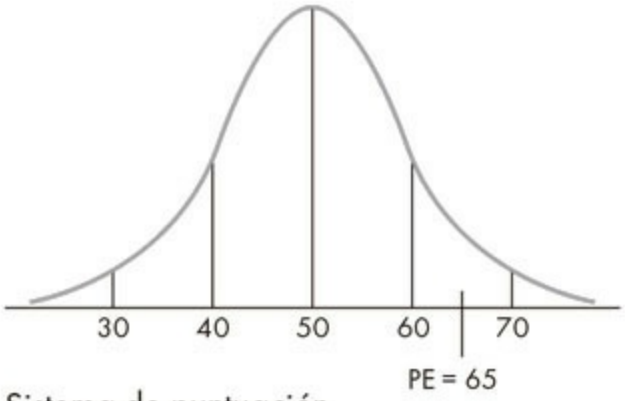
En la práctica común, todos estos pasos ya están hechos, por lo que basta con usar los cuadros del manual de la prueba para convertir una puntuación natural en estándar. La figura 3-13 ofrece un ejemplo de estos cuadros; incluye sólo una sección del cuadro, no todo el rango de puntuaciones.



Puntuaciones naturales
M = 38
DE = 6
X = 47



Puntuaciones z
M = 0.0
DE = 1.0
z = 1.5



Puntuaciones estándar
M = 50
DE = 10
PE = 65

Sistema de puntuación natural $M = 38 \quad DE = 6 \quad X = 47$

Puntuación z $z = (X - M)/DE = (47 - 38)/6$

Puntuaciones estándar $M = 50 \quad DE = 10 \quad PE = 65$

Fórmula 3–5: $PE = (DE_e / DE_n)(X - M)_n + M_e$

$PE = (10/6)(47 - 38) + 50 = 65$

Figura 3-12. Conversión del sistema de puntuación natural al de puntuación estándar.

Puntuación natural:	...	60	61	62	63	64	65	66	67	68	69	...
Puntuación estándar:	...	55	56	56	57	58	59	60	61	62	63	...

Figura 3-13. Ejemplo de una conversión de puntuación natural en estándar (puntuación T).

Lineal o no lineal

La mayoría de las puntuaciones estándar son **transformaciones lineales** de las puntuaciones naturales y se obtienen mediante la fórmula 3-5. Sin embargo, algunas se derivan de una **transformación no lineal**, en cuyo caso la fórmula 3-5 y el ejemplo de la figura 3-12 no se pueden aplicar. Las transformaciones no lineales pueden emplearse para obtener una distribución de puntuaciones que sea normal. Por ello, a veces nos referimos al resultado como **puntuación estándar normalizada**; el efecto de la transformación no lineal se muestra en la figura 3-14. Por lo general, la transformación no lineal se realiza con base en la relación entre puntuaciones z y percentiles en las áreas bajo la curva normal, por lo que una transformación no lineal a veces se denomina área de transformación. Aunque suena complicado, en realidad es bastante sencillo. A lo largo de este capítulo asumimos que las puntuaciones estándar son transformaciones lineales a menos que se indique lo contrario.

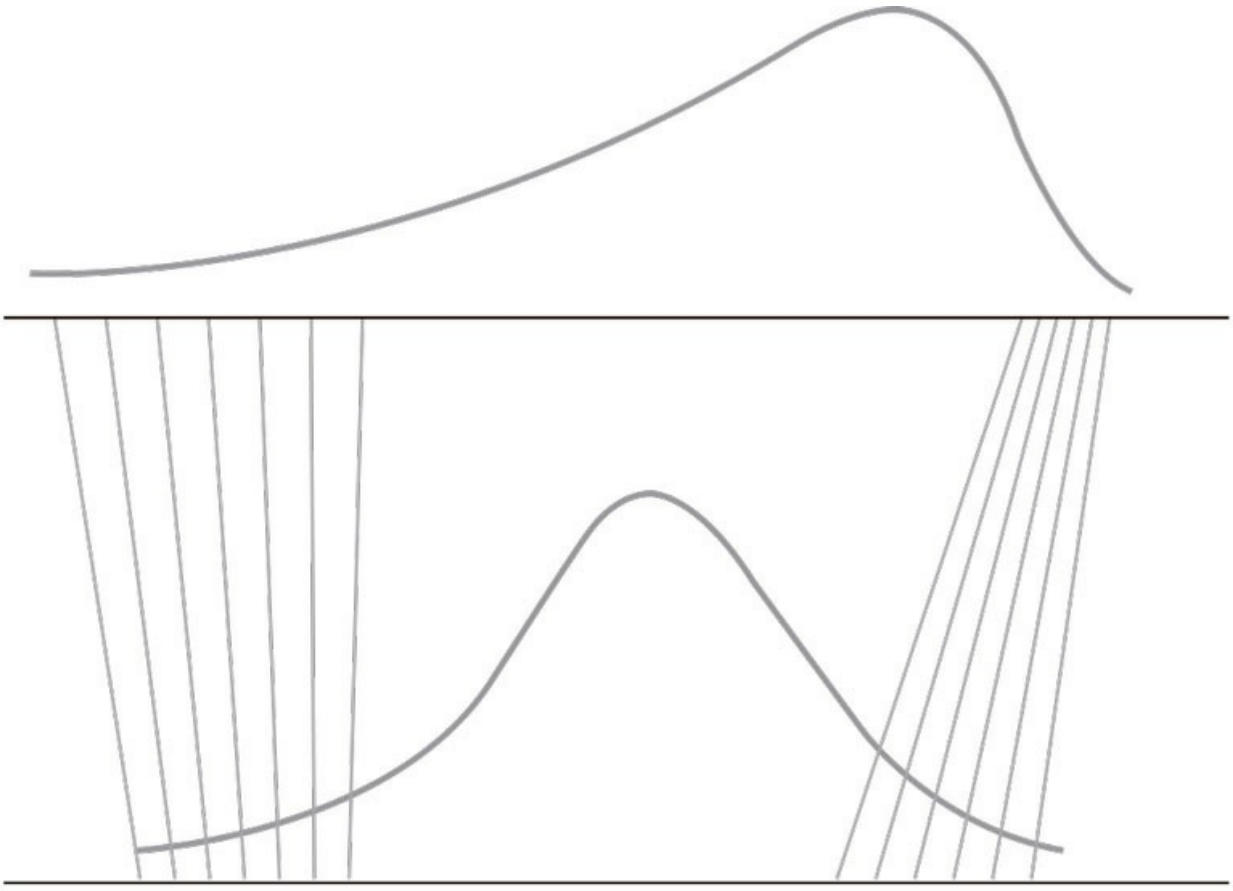


Figura 3-14. Ejemplo de transformación no lineal de una distribución no normal de puntuaciones naturales en un sistema de puntuación estándar que se aproxima a la distribución normal.

Resumen de puntos clave 3-4

Algunos tipos de puntuaciones estándar
Puntuaciones T
SAT y GRE
CI de desviación
Estaninas o eneatipos
Equivalentes de la curva normal
Puntuaciones estándar (o escalares) de multinivel

Puntuaciones T

Las **puntuaciones T**, a veces llamadas puntuaciones T de McCall, son puntuaciones estándar con $M = 50$ y $DE = 10$. Así, el rango real de estas puntuaciones va de 20 (que corresponde a −3z) a cerca de 80 (+3z). Las puntuaciones T (escrito con T mayúscula) deben distinguirse de los valores t (con t minúscula) de Student que se usan en las

pruebas de significancia estadística. Las puntuaciones T son muy utilizadas en pruebas de personalidad, aunque también se usan en pruebas de otro tipo. Por ejemplo, el **Inventario Multifásico de Personalidad de Minnesota** (MMPI) y el **Strong Interest Inventory** (SII), que se describen en los capítulos 13 y 15, respectivamente, emplean puntuaciones T. Las figuras 3-10a y 3-10b muestran la distribución de las puntuaciones T.

SAT y GRE

El SAT (antiguamente *Scholastic Assessment Test*) utiliza un sistema de puntuaciones estándar con $M = 500$ y $DE = 100$. Este sistema se aplica a las principales pruebas de las series: Lectura crítica (antes Verbal), Matemáticas y Escritura. Las pruebas, a menudo, se combinan para obtener una puntuación total; por ejemplo, la puntuación combinada de Verbal y Matemáticas a menudo se usa en el SAT. Cuando esto se hace, las medias son aditivas, pero las *DE* no; es decir, la media de la puntuación combinada o Total es 500 + 500, pero la *DE* de la puntuación total no es 100 + 100, sino que es menor de 200, ya que las dos pruebas que se combinan no están perfectamente correlacionadas. Este fenómeno no es exclusivo del SAT, sino que ocurre en cualquier combinación de pruebas cuya correlación no sea perfecta.

Hasta agosto de 2011, el GRE General Exam también usaba un sistema de puntuaciones estándar con $M = 500$ y $DE = 100$ (y un rango de 200 a 800) en sus pruebas de Razonamiento verbal y Razonamiento cuantitativo. Después de esa fecha, el GRE cambió a un sistema con un rango de puntuaciones de 120 a 170. Aunque no se clasifica oficialmente como tal, al inspeccionar los cuadros de normas se observa que el sistema de puntuación subyacente tiene, aproximadamente, una $M = 150$ y $DE = 10$. El GRE Subject Exams sigue usando el viejo sistema de puntuación estándar ($M = 500$, $DE = 100$).

Aquí, al usar las normas del SAT y GRE, surge un asunto muy confuso, que puede aparecer con las normas de cualquier prueba, pero parece especialmente agudo con las puntuaciones del SAT y GRE. Recordemos nuestra descripción de las figuras 3-10a y 3-10b; en un sistema de puntuación estándar con $M = 500$ y $DE = 100$, una puntuación de 500 debe ubicarse en el percentil 50, una puntuación de 600, en el percentil 84, una puntuación de 400, en el percentil 16, y así sucesivamente. La inspección de los cuadros de normas del SAT y GRE muestra que estas relaciones no se mantienen; de hecho, a menudo están lejos de estas relaciones, para consternación de los examinados. Por ejemplo, a partir de 2012, una puntuación Verbal de GRE de 500 se ubica en el percentil 62, mientras que una puntuación Cuantitativa de GRE de 500 se encuentra en el percentil 26. Dicho de otro modo, para estar aproximadamente en el percentil 50, se necesita una puntuación de 460 en Verbal, pero una de 620 en Cuantitativa. ¿Cómo puede suceder esto? La explicación es muy sencilla. El sistema de puntuación estándar con $M = 500$ y $DE = 100$ se establece en un punto en el tiempo, pero las normas de percentiles se ajustan cada año con base en las personas que han respondido la prueba en los tres años

más recientes. Así, las normas de percentiles pueden "desfasarse" conforme las capacidades de los examinados cambian en el curso de los años. Mientras tanto, el sistema con $M = 500$ y $DE = 100$ se mantiene fijo. Cada cierto periodo de tiempo, la editorial de la prueba reajusta el grupo base para el sistema de puntuación estándar, realineando así los sistemas de puntuación estándar y de percentiles. Entonces, continúa el desfase. Este confuso asunto confirma la importancia de saber sobre las normas y su origen para poder interpretar de una manera adecuada las puntuaciones de las pruebas.

CI de desviación

La definición tradicional de CI (**cociente de inteligencia**) es: CI = (EM/EC) × 100, donde EM significa **edad mental** (véase la página 61a» para ver la descripción de este término), EC, edad cronológica, y 100 es un multiplicador para eliminar el punto decimal. Por ejemplo, la edad mental de Olivia es de 10 años y su edad cronológica es de 8 años; por lo tanto, su CI es (10/8) × 100 = 125. Esto es un **CI de razón**, ya que representa la razón entre EM y EC.

¡Inténtalo!

Calcula los CI de razón de los siguientes casos.

La EM de Matt es de 78 meses (6 años, 6 meses) y su EC es de 84 meses (7 años, 0 meses). ¿Cuál es su CI de razón?

La EM de Meg es de 192 meses y su EC es de 124 meses. ¿Cuál es su CI de razón?

Los CI de razón se usaron en las primeras pruebas de inteligencia; sin embargo, se observó que las desviaciones estándar de estos CI no eran iguales en distintos niveles de edad. En particular, las desviaciones estándar tendían a aumentar con la edad. Así, el CI de razón de 120 se desvía menos del promedio (100) a la edad de 18 que a la de 6. A la inversa, un CI de razón de 70 se desvía más de 100 a la edad de 7 años que a la de 17. Tal variación en el significado de un CI en distintos niveles de edad es desafortunada e indeseable.

Los CI que se obtienen de las modernas pruebas de inteligencia *no* son de razón, sino que son puntuaciones estándar con $M = 100$ y *generalmente* DE de 15 o 16. Estas puntuaciones estándar a menudo se denominan **CI de desviación**. La $M = 100$ se usa en defensa de la tradicional (razón) definición de CI. Los CI de razón de la prueba original Stanford-Binet proporcionaban una desviación estándar de 16 en ciertas edades, por lo que ésta se adoptó como *la DE* de las puntuaciones estándar usadas en algunas pruebas de inteligencia. Otras pruebas, entre las cuales las más notables son las pruebas Wechsler (WAIS, WISC, WPPSI), adoptaron la $DE = 15$.

Algunos sectores de la comunidad psicológica se esfuerzan con afán para rechazar el término CI, mientras que otros mantienen la tradición de usar un sistema de puntuación estándar con $M = 100$ y $DE = 15$ o 16. De ahí que a lo que nosotros nos referimos como

CI de desviación a veces sale a la superficie en los manuales de las pruebas y en los informes de puntuaciones bajo diferentes nombres, por ejemplo, school ability index (SAI [índice de capacidad escolar]). Por lo general, estos nombres alternos pueden reconocerse con facilidad como sistemas de puntuación estándar con una *M* y *DE* conocidas.

Estaninas

Las **estaninas,** contracción de "standard nine" [nueve estándar], son un sistema de puntuación estándar con $M = 5$ y $DE = 2$ (aproximadamente). Las estaninas se crearon para a) dividir la distribución normal en nueve unidades y b) tener unidades que cubrieran distancias iguales en la base de la curva normal, excepto por las unidades que cubren las colas de la distribución, es decir, las unidades 1 y 9. Cuando se cumplen estas condiciones, la media será obviamente 5 y la desviación estándar será ligeramente mayor de 2. Estas dos propiedades de las estaninas se ilustran en la figura 3-10a; véase también la figura 3-15. Podemos notar que las unidades 2 a la 8 cubren distancias iguales en la base de la curva normal. Ya que la densidad de la curva varía en diferentes secciones, estas distancias iguales contienen porcentajes variables de casos en la distribución; por ejemplo, la estanina 2 contiene el 7% de los casos (del percentil 4 al 11) mientras que la estanina 4 contiene el 17% de los casos (del percentil 23 al 40).

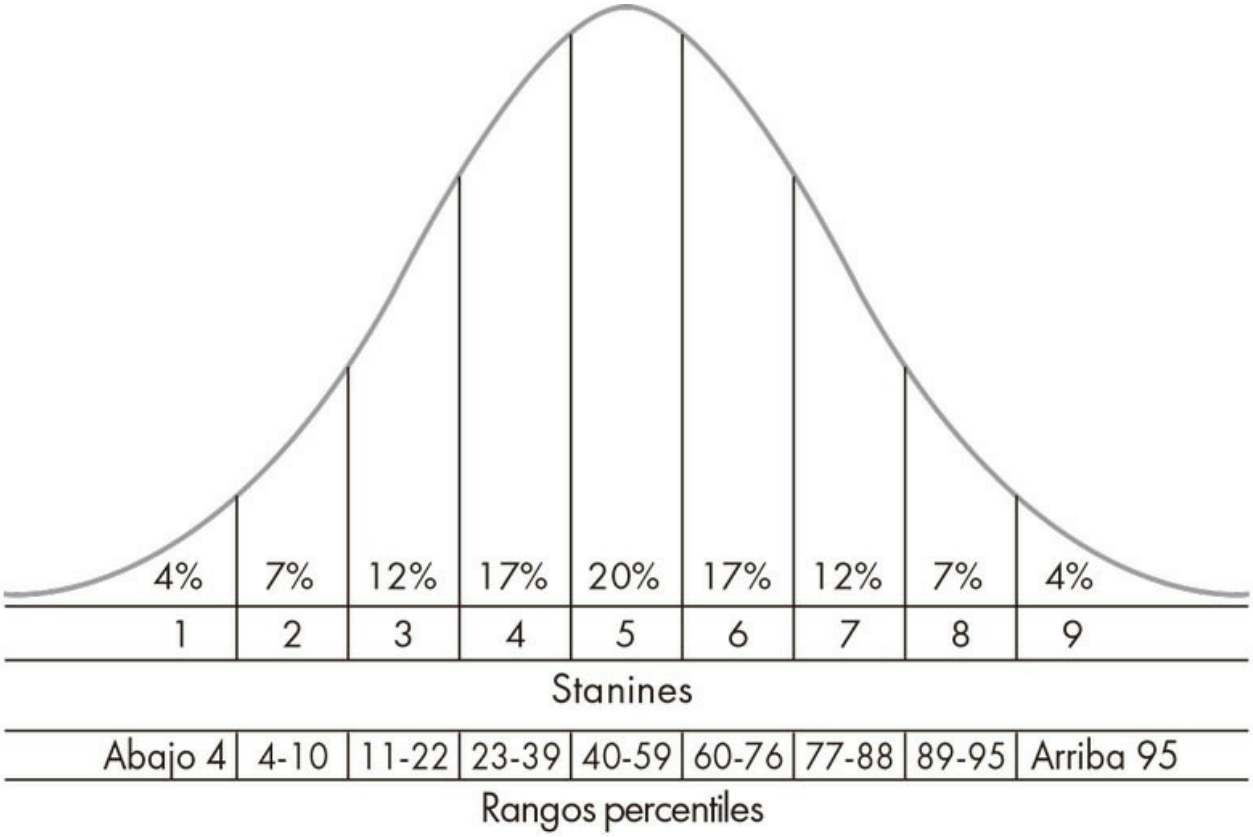


Figura 3-15. Distribución de estaninas.

¡Inténtalo!

Usando la figura 3-10 o 3-15, determina la estanina que corresponde a los siguientes percentiles:
Percentil: 2 36 25 90
Estanina: _____ _____ _____ _____

Las estaninas siempre se derivan por referencia a las divisiones de percentiles que se muestran en la figura 3-15 en vez de la fórmula 3-5. De ahí que las estaninas resultan en una transformación no lineal de puntuaciones naturales (a menos que la distribución original haya sido perfectamente normal). Se usan con mucha frecuencia para hacer informes de las puntuaciones de pruebas estandarizadas de aprovechamiento y de algunas pruebas de capacidad mental en escuelas primarias y secundarias, pero no se emplean mucho en otros contextos.

Equivalentes de la curva normal

El **equivalente de la curva normal** (ECN) es un sistema de puntuación estándar creado de modo que los ECN sean iguales a los rangos percentiles en los puntos 1.50 y 99. Cuando se cumple esta condición, el sistema ECN tiene una $M = 50$ y $DE = 21$ (aproximadamente). Los ECN se usan casi exclusivamente para cumplir con ciertos requerimientos para informes federales sobre las pruebas de aprovechamiento en escuelas públicas. La figura 3-10a muestra la relación entre los ECN y otras normas.

Puntuaciones estándar de multinivel

Una prueba de multinivel tiene, al menos de manera parcial, distintas pruebas en diferentes niveles de edad o grado. Los ejemplos primarios de estas pruebas son las baterías de aprovechamiento (véase capítulo 11) y las pruebas de capacidades cognitivas de aplicación grupal (véase capítulo 9) que se usan en las escuelas primarias y secundarias. Aunque estas pruebas pueden tener el mismo nombre (p. ej., *Metropolitan Achievement Tests* u *Otis-Lennon School Ability Test*) a lo largo de un amplio rango de edad o grados escolares, es evidente que no se usan los mismos reactivos en todo el rango. La prueba se divide en varios niveles separados; un nivel puede usarse para los grados 1 y 2, otro, para los grados 3 y 4, y así sucesivamente.

Las puntuaciones naturales obtenidas en distintos niveles de estas pruebas a menudo están ligadas por un sistema de puntuaciones estándar que abarcan todos los niveles. Estas puntuaciones estándar se denominan, a veces, **puntuaciones en la escala**. Las puntuaciones estándar de multinivel son difíciles de interpretar. Tienen una media y una desviación estándar cómodas (es decir, 500 y 100) en un nivel, que suele ser el grado o

edad que se encuentra a la mitad del rango total; sin embargo, tienden a diferir en las otras edades o grados. De este modo, una puntuación estándar de 673 en una prueba de lectura de un estudiante de séptimo grado no tiene un significado interpretable sencillo.

Las puntuaciones estándar de multinivel pueden ser útiles para medir el crecimiento a lo largo de los grados o edades. Este sistema de puntuación estándar se elabora para aproximarse a una escala intervalar. Sin embargo, para la interpretación ordinaria de la prueba, estas puntuaciones no son muy útiles.

Fortalezas y debilidades de las puntuaciones estándar

Las puntuaciones estándar brindan una medida práctica para interpretar el desempeño en una prueba en distintas circunstancias. Ya que muchos rasgos de interés para los psicólogos, supuestamente tienen una distribución normal, las conexiones de las puntuaciones estándar con las puntuaciones z son útiles. Las puntuaciones estándar evitan el problema de los percentiles relacionado con la desigualdad de sus unidades en varias regiones de la distribución normal. Por esta razón, las puntuaciones estándar son más flexibles para los cálculos estadísticos.

Las puntuaciones estándar tienen algunos inconvenientes. Primero, debe admitirse que una fracción extremadamente pequeña de la raza humana tiene idea de qué es una curva normal o una puntuación z. De ahí que relacionar puntuaciones estándar con el contexto de la curva normal y las puntuaciones z tenga poco valor, excepto cuando se trabaja con expertos. Segundo, para darle significado a una puntuación estándar, necesitamos tener presentes la *M* y *DE* de ese sistema. En párrafos anteriores citamos algunos de los sistemas más conocidos de puntuación estándar en los que este problema se minimiza, por ejemplo, $M = 100$ y $DE = 15$ en el caso de los CI de desviación en las pruebas de capacidad mental. Sin embargo, existen muchos otros sistemas de puntuación estándar: una variedad potencialmente infinita, ya que se puede escoger cualquier valor para *M* y *DE*. Por ejemplo, el *Law School Admissions Test* (LSAT) y el ACT College Entrance Test tienen, cada uno, su sistema de puntuación estándar distintivo. ¿Qué significa una media de 130 en el LSAT? ¿Qué significa una puntuación de 26 en el ACT? No se puede tener idea sin consultar en el manual la *M* y *DE* de sus sistemas.

Las estaninas merecen un comentario especial. Tienen la virtud de su sencillez para informar las puntuaciones individuales; es sencillo explicar, por ejemplo, a los padres que el desempeño de su hijo corresponde a una cantidad en una escala del 1 al 9. En general, no se necesita ninguna explicación adicional acerca de las medias, desviaciones estándar, distancias iguales en la base de la curva normal, etc. Esta sencillez es una ventaja. Por otro lado, las estaninas son más bien burdas para informar los promedios grupales.

Los equivalentes de la curva normal (ECN) son un invento desafortunado. Como puntuaciones estándar no ofrecen ninguna ventaja sobre otros sistemas; sin embargo, parecen percentiles y, por eso, se confunden fácilmente con ellos. Como señalamos antes, los ECN coinciden con los percentiles en tres puntos de la distribución normal, pero son muy diferentes en otros puntos.

Normas de desarrollo

Cuando el rasgo que se mide se desarrolla de manera sistemática en el tiempo, es factible crear lo que se denomina una **norma de desarrollo**. Existen dos normas de este tipo que son muy usadas: **equivalentes de edad** (EE) y **equivalentes de grado** (EG). Los EE se usan en algunas pruebas de capacidad mental, en cuyo caso la puntuación se denomina edad mental (EM), por mucho la más conocida de los equivalentes de edad. Los EG se usan en muchas pruebas de aprovechamiento. Las normas de desarrollo sólo tienen sentido en la medida en que el rasgo que se mide se desarrolla o crece con el tiempo en la población pertinente. En una norma de desarrollo, una puntuación natural se interpreta en términos de la edad o grado para el que dicha puntuación es típica.

Edad mental (EM) «61a

La edad mental es el ejemplo principal de los equivalentes de edad. La EM fue uno de los primeros tipos de normas que se usaron en las pruebas psicológicas. Tuvo su origen en las escalas de Binet. La EM se determina encontrando la puntuación típica o media de los examinados en niveles de edad sucesivos. Los grupos de edad pueden formarse por intervalos de un año, medio año, tres meses o cualquier otra forma semejante de agrupar individuos. Luego se determina la puntuación mediana en las pruebas para cada grupo. Los resultados se grafican y una suave curva se adecua a los puntos como en la figura 3-16. Cada "•" que aparece en la figura es una mediana obtenida del proceso de obtención de normas de la prueba. La figura ilustra cómo obtener la EM a partir de una puntuación natural. Por ejemplo, un niño que obtiene una puntuación natural de 42 tiene una EM de 88 meses o 7 años 4 meses (a menudo, escrito como 7–4 en el lenguaje de los equivalentes de edad). En la práctica, las puntuaciones naturales se convierten en edad mental con ayuda de un cuadro preparado a partir de la curva de desarrollo; la figura 3-17 es un ejemplo de tal cuadro.

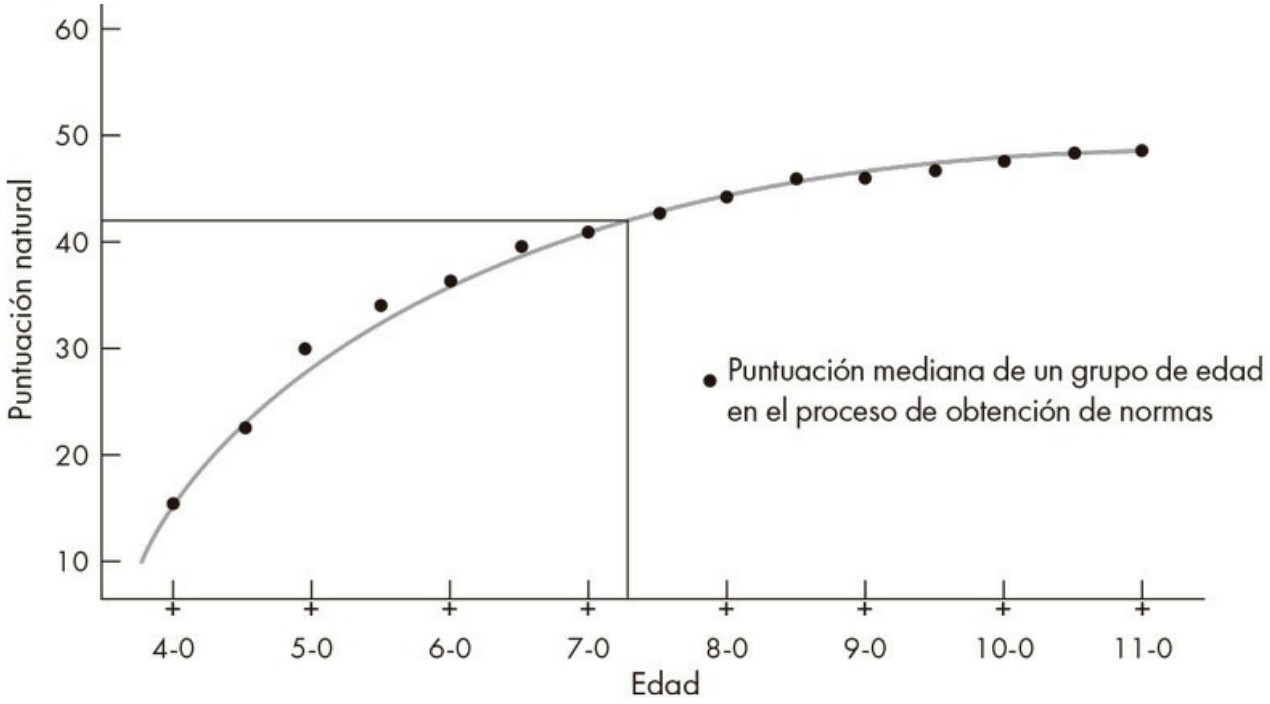


Figura 3-16. Ilustración de la curva para desarrollar las edades mentales.

Equivalentes de grado (EG)

Puntuación natural	15	20	25	30	35	40	45
Edad mental	4-0	4-2	4-5	5-0	5-4	6-2	8-8

Figura 3-17. Cuadro de correspondencias entre puntuación natural y edad mental creado a partir de la figura 3-16.

Los equivalentes de grado se elaboran aplicando una prueba a estudiantes de diferentes grados, lo cual se hace en el proceso de obtención de normas. Se obtiene el desempeño típico o mediano de cada grado. Los puntos medianos se grafican y una curva se adecua a los puntos, como en la figura 3-16, pero con grados en lugar de edades en la base. Semejante al procedimiento de la EM, la EG correspondiente a una puntuación natural se lee a partir de la curva, y se prepara un cuadro para las conversiones de puntuación natural en EM.

Lo convencional para los EG es dividir el año escolar en 10 partes, como se muestra en la figura 3-18. Un EG se informa, por ejemplo, como 6.3, lo que significa tercer mes del grado 6. A menudo se reportan los EG arriba de 12.9 (el último mes del grado 12) como

12.9+ o con alguna etiqueta verbal, como Post High School (PHS). La escala de EG no suele extenderse a los años universitarios.

Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun
0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

Figura 3-18. División del año escolar en el sistema de equivalentes de grado.

Otras normas de desarrollo

Aunque la edad mental y los equivalentes de grado son los principales ejemplos de las normas de desarrollo, deben mencionarse brevemente otros dos ejemplos. Primero, hay pruebas basadas en las *teorías de las etapas* del desarrollo humano. Ejemplos bien conocidos son la teoría de Piaget del desarrollo cognitivo y la de Kohlberg del desarrollo moral. Las pruebas basadas en estas teorías proporcionan resultados que ubican al individuo en cierta etapa. Por ejemplo, una tarea piagetiana puede ubicar a un niño en la "etapa preoperacional".

Un segundo ejemplo son las mediciones antropométricas, como el peso y la estatura. A menudo, tales mediciones se interpretan en términos de normas de desarrollo, que son, en esencia, equivalentes de edad. Por ejemplo, se informa que un niño es "tan alto como un niño promedio de 6 años de edad". Justo como la edad mental, tales afirmaciones se suelen interpretar en relación con la edad cronológica del niño, por ejemplo, "Miguel es muy alto para su edad".

Fortalezas y debilidades de las normas de desarrollo

Todas las normas de desarrollo tienen algunas fortalezas y debilidades en común. Del lado positivo, tienen una naturalidad en su significado que las hace muy atractivas. Decir que una persona de 16 años de edad funciona mentalmente como una de 3 o que un estudiante de segundo grado puede leer al nivel de uno de octavo son afirmaciones que parecen transmitir un significado considerable, libre de la jerga estadística estéril de los rangos percentiles y las puntuaciones estándar. Las ideas sobre los patrones de desarrollo normal están profundamente arraigadas en nuestra forma de pensar sobre los humanos. La noción básica de las normas de desarrollo se usa en muchas situaciones. El adulto que hace un berrinche es acusado de "actuar como un niño de 2 años". Se puede observar que los estudiantes de sexto grado en Japón estudian el álgebra "que por lo general se reserva para los alumnos de noveno grado en EUA". Miguel recibe elogios porque "se

desempeñó en su primer juego como si tuviera mucha experiencia". Los equivalentes de edad y de grado simplemente formalizan estas maneras de pensar y ayudan a cumplir la meta de dar sentido a las puntuaciones naturales.

La segunda ventaja de las normas de desarrollo es que proporcionan una base para medir el crecimiento en las pruebas multinivel. Por ejemplo, en los años de escuela elemental, un niño puede responder el nivel I de primaria de una prueba de aprovechamiento en primer grado, el nivel elemental en cuarto grado y el nivel intermedio en séptimo grado. Los equivalentes de grado vinculan todos estos niveles de la prueba.

Las normas de desarrollo tienen dos principales inconvenientes. Primero, sólo se pueden aplicar a variables que muestren patrones claros de desarrollo. De ahí que no puedan aplicarse, por lo general, en áreas como rasgos de personalidad, actitudes e intereses vocacionales. Por ejemplo, no significa nada decir que alguien tiene la extroversión de un niño de 10 años o de un alumno de tercer grado. Además, incluso las variables que tienen patrones de desarrollo en algún nivel no continúan de manera ordinaria con sus patrones de crecimiento indefinidamente. Por ejemplo, la capacidad mental, tal como se mide con las pruebas, se desarrolla de modo sistemático hasta la edad aproximada de 18 años, pero no después. La capacidad de lectura se desarrolla con rapidez en la escuela primaria, pero no sigue desarrollándose así indefinidamente. Hay una distinción útil para interpretar las pruebas entre la capacidad mental de un niño de 5 año y un joven de 15, pero no entre uno de 25 y otro de 35 años. Ésta no es una cuestión de todo o nada; las normas de desarrollo no dejan de ser útiles en un punto definido con claridad, sino que lo hacen de manera gradual conforme la curva de desarrollo se vuelve menos empinada (véase arriba la figura 3-16). Cuando la curva se vuelve plana, las normas de desarrollo pierden por completo su sentido.

La segunda desventaja de las normas de desarrollo son sus descontroladas desviaciones estándar. Las *DE* suelen ser diferentes en distintos niveles o distintas pruebas, pues varían de manera no sistemática entre distintas pruebas de la misma batería, incluso cuando las normas se basan en el mismo grupo de normalización. Aunque esto pueda sonar trivial o demasiado técnico, tiene implicaciones prácticas importantes. Por ejemplo, un niño de 5 años de edad que se encuentre un año por debajo del promedio en capacidad mental (es decir, EC = 5–0, EM = 4–0) está mucho más abajo del promedio en unidades de desviación estándar o de percentiles que un joven de 16 años que se encuentra un año por debajo del promedio (es decir, EC = 16–0, EM = 15–0). Un estudiante del grado 1.5 que lee al nivel EG = 3.5 está prácticamente fuera de la distribución de los estudiantes del grado 1, mientras que uno del grado 7.5 que lee al nivel de EG = 9.5 no es tan extraordinario. Consideremos también al estudiante del grado 3.5 con un EG = 4.5 tanto en cálculos matemáticos como en lectura; en comparación con otros estudiantes, éste se encuentra probablemente mucho más avanzado en cálculo que en lectura porque la *DE* de los GE es, por lo general, menor en cálculo que en lectura. Estas diferencias en las *DE* de varias pruebas no son sistemáticas, sino que pueden diferir de una serie de pruebas a otra.

La tercera crítica que suele hacerse a los equivalentes de grado es la siguiente: se ha observado que un estudiante, digamos, del grado 3 puede obtener un EG de 6.5 sin saber el material de un estudiante típico del grado 6, pues basta que responda de manera perfecta todos los reactivos de los grados 2, 3 y 4, mientras que el estudiante típico del grado 6 tiene correctos sólo algunos reactivos, no todos, de los grados 2 al 7. Esta situación se describe en la figura 3-19.

Nivel de contenido	Grado 2	Grado 3	Grado 4	Grado 5	Grado 6	Grado 7	PN	EG
Reactivos	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx		
Estudiante del grado 3	/////	/////	/////				15	6.5
Estudiante del grado 6	////	///	///	///	/	/	15	6.5

Cada x representa un reactivo de la prueba y cada /, una respuesta correcta.

Figura 3-19. Diferentes maneras de obtener un EG de 6.5.

Este argumento, que se suele mencionar sólo en relación con los equivalentes de grado, puede aplicarse a cualquier tipo de puntuación normativa. La puntuación de dos estudiantes ubicados en el percentil 75 o en una puntuación estándar de 60 no necesariamente implica que tuvieron correctos los mismos reactivos. Las pruebas elaboradas de acuerdo con la teoría de la respuesta al reactivo intentan minimizar este problema, pero éste es una función del método de elaboración de la prueba, no del tipo de normas empleadas.

Ejemplos de cuadros de normas

Ahora que hemos considerado los principales tipos de normas, será de utilidad observar el modo en que aparecen en los manuales de las pruebas. ¿Qué apariencia tienen los cuadros de normas? En realidad, estos cuadros aparecen en una gran variedad de formas; sin embargo, hay algunos patrones estándar. Después de ver algunos ejemplos, es fácil descifrar las variaciones de estos patrones.

El cuadro 3-3 muestra la presentación típica de un cuadro de normas. Siempre se empieza con la puntuación natural (PN), que luego se convierte en puntuación normativa. En este ejemplo, primero se busca una puntuación natural en la columna de la izquierda y luego alguna de las diversas puntuaciones escalares: puntuación estándar (PE), CI de desviación (CID), rango percentil (RP), estaninas (S) y equivalente de curva normal (ECN). En muchas aplicaciones se realiza la conversión con ayuda de una computadora, y lo único que se ve es la puntuación escalar en el informe. En la práctica, solemos concentrarnos en sólo un tipo de puntuación normativa aun cuando citamos varias.

Cuadro 3-3. Ejemplo de un cuadro de normas con varios tipos de normas

PN	PE	CID	RP	S	ECN
60	420	119	90	8	77
59	417	117	86	7	74
58	414	114	82	7	69

Informes interpretativos y normas

Recordemos que el propósito fundamental de las normas es ofrecer un contexto para interpretar una puntuación natural. Por lo general, la información normativa es cuantitativa: otro conjunto de números. Sin embargo, las puntuaciones de las pruebas cada vez se informan con mayor frecuencia en forma de narrativa generada por computadora. El usuario puede no ver ningún número, pero la mayoría de los informes proporcionan tanto números –las normas usuales– como un **informe interpretativo**. ¿Cómo se originan estos informes?

La sustancia de los informes interpretativos siempre empieza con la puntuación de una prueba, al menos una puntuación natural o theta, y casi siempre con una puntuación escalar. Desde el principio, el informe varía considerablemente en complejidad; en el nivel más sencillo, puede sólo traducir una puntuación normativa a una descripción verbal. Por ejemplo, una computadora puede tener un cuadro que muestra la siguiente correspondencia entre puntuaciones estándar (en un sistema con $M = 100$, $DE = 15$) y categorías verbales:

130+	Excepcionalmente alto
120 – 129	Muy arriba del promedio
110 – 119	Arriba del promedio
90 – 109	Promedio
80 – 89	Abajo del promedio
70 – 79	Muy abajo del promedio
Menos de 70	Excepcionalmente bajo

Con este cuadro, el perfil de puntuaciones que obtuvo un individuo en las pruebas A, B y C puede verse así:

Prueba	Puntuación	Nivel de desempeño
A	98	Promedio
B	82	Debajo de la media
C	94	Promedio

La columna “Puntuación” no puede aparecer en el informe aunque las puntuaciones

sean su punto de partida. Con un programa de cómputo un poco más sofisticado, el informe podría leerse así: "El desempeño de Luis en las pruebas A y C se ubicó en el rango promedio, mientras que su desempeño en la prueba B estuvo un poco por debajo del promedio". Los reportes interpretativos a menudo hacen referencia al grupo normativo, por ejemplo, "En comparación con otros muchachos de su grado, Luis se encuentra en el percentil 60 en aptitud mecánica, es decir, está ligeramente arriba del promedio de los muchachos de su grado".

Algunos informes interpretativos van mucho más allá de la traducción de puntuaciones escalares a categorías verbales, pues toman en cuenta información sobre la confiabilidad y validez de la prueba y realizan múltiples comparaciones entre puntuaciones. Algunos informes ocupan una docena de páginas a espacio seguido.

¡Inténtalo!

Inventa un conjunto de categorías que distingan entre las estaninas 1-3, 4-6 y 7-9 para usarlas en un informe interpretativo. Las categorías pueden ser de más de una palabra.

Grupo de tanines	Categoría verbal
1-3	_____
4-6	_____
7-9	_____

La figura 3-20 muestra un informe interpretativo simulado de resultados de pruebas aplicadas a un estudiante universitario recién ingresado al primer año. Podemos notar que el informe nunca emplea números aunque éstos (primero las puntuaciones naturales, luego sus percentiles equivalentes) sean el punto de partida. Aquí señalamos algunos puntos sobre cómo se puede elaborar dicho informe con un programa de cómputo relativamente sencillo. El primer enunciado tiene una redacción común en la que se indica el mes y los nombres de las pruebas aplicadas. La segunda oración es común a todos los informes. La tercera oración parte de un cuadro como el que acabamos de presentar y, luego, se convierten los percentiles en palabras: "arriba del promedio" se refiere a percentiles arriba de 70 y "muy alto", a percentiles arriba de 90. Los siguientes dos enunciados se basan en estudios sobre la validez de las pruebas –lo que llamamos validez predictiva en el capítulo 5. ¡El último enunciado del primer párrafo se aplica a todos! Las afirmaciones sobre el curso en que debe ubicarse al examinado, que se presentan en el segundo párrafo, se basan en los juicios del profesorado sobre los niveles de las puntuaciones requeridas para dichos cursos. El programa de cómputo jala el último enunciado de este párrafo para cualquiera que se ubique en el curso de español 201 o superior. El tercer párrafo es común a todos los reportes.

Asesoría Académica Universitaria
20 de junio de 2013

Querida Alicia,

A principios de junio hiciste los exámenes universitarios de colocación en escritura, matemáticas y lengua extranjera, y también respondiste preguntas sobre tus intereses académicos. Los resultados de estos exámenes ayudarán a planear tu primer año en la Universidad. Tu desempeño estuvo por encima del promedio en escritura y matemáticas, de hecho, muy alto en el examen de matemáticas. Los estudiantes de tu nivel suelen tener éxito en sus estudios en la Universidad. A menudo continúan sus estudios de posgrado o especialización con éxito, lo cual es importante si consideramos lo que indicaste sobre tus intereses académicos. Desde luego, ¡tienes que trabajar duro para hacerlo bien!
En vista de tu desempeño en matemáticas, te recomendamos tomar el curso 150 de matemáticas en tu primer semestre. También te recomendamos el curso de 140 de inglés. En cuanto a la lengua extranjera (hiciste el examen de colocación de español), puedes omitir los cursos 101 y 102 de español y entrar directamente al curso 201. Ese será un buen inicio para ti e, incluso, te permitirá considerar el español como un curso secundario.
Bienvenida a la Universidad. Esperamos verte en el semestre de otoño. La mejor de las suertes. Si deseas discutir tu desempeño en los exámenes de colocación, por favor, solicítalo poniéndote en contacto con la oficina de Asesoría Académica al teléfono 700-111-2222 o en el correo electrónico AcadAdvis@daU.edu.

Dr. Cordial
Director de Asesoría Académica

Figura 3-20. Ejemplo de un informe interpretativo de un programa de examinación. «65a

¡Inténtalo!

En años recientes se ha desarrollado una auténtica industria artesanal psicométrica para producir informes interpretativos de resultados de pruebas. Puedes encontrar ejemplos de tales informes en las páginas web de las editoriales. Usa uno de los siguientes sitios para examinar uno o dos ejemplos de informes sobre pruebas muy utilizadas:
http://psychcorp.pearsonassessments.com/pai/ca/research/publications/samplerpts/reslist.htm
http://www4.parinc.com/WebUploads/samplerpts/PAI32.pdf
http://portal.wpspublish.com/portal/page?_pageid=53,124601&_dad=portal&_schema=PORTAL

Efecto Barnum

Al evaluar los informes intepretativos de las puntuaciones de las pruebas, debemos estar particularmente atentos al funcionamiento del efecto Barnum. Este fenómeno se denomina así por el promotor de circo P. T. Barnum, famoso por su capacidad de hacer creer (y comprar) a la gente cualquier cosa. Cuando se aplica al campo de las pruebas psicológicas, el **efecto Barnum** se refiere a la tendencia de la gente a creer afirmaciones rimbombantes que probablemente sean ciertas para todo mundo y que no contienen información específica y única proveniente de la prueba. Consideremos estas afirmaciones supuestamente basadas en una prueba de personalidad:

- Tus puntuaciones indican que, con algunos grupos, puedes ser muy extrovertido. Sin embargo, a veces quieres estar solo.
- De acuerdo con esta prueba, hay veces en que piensas que no es justo que la gente se aproveche de otras personas.

Ahora consideremos estas afirmaciones basadas en pruebas de capacidad mental y de rendimiento aplicadas a José y Abigail, estudiantes de escuela primaria:

- La batería de pruebas muestra que José no es igual de bueno en todas las áreas. Su maestro necesita sacar provecho de sus fortalezas y motivarlo para mejorar en otras áreas.
- Estos resultados pueden ser de especial ayuda para tratar casos como el de Abigail. Sin duda, ella puede mejorar de muchas maneras.

Todas estas afirmaciones podrían hacerse sobre casi cualquier persona, pues no contienen información específica basada en los resultados de las pruebas. De hecho, podríamos escribir un informe entero basado en estas afirmaciones estilo Barnum, las cuales no son útiles. Los informes narrativos deben dar información que sólo caracterice al individuo y surja directamente de los resultados de las pruebas.

¡Inténtalo!

Escribe afirmaciones estilo Barnum en el campo de la personalidad. Deben poder aplicarse a casi cualquier persona que conozcas. Empieza con algo así: Los resultados de la prueba muestran que tú...

Grupos normativos

Todos los tipos de normas que tratamos antes en este capítulo se basan en los grupos normativos. La prueba se aplica a un grupo como parte de lo que se denomina *programa de obtención de normas* o *de estandarización*. El valor de las normas de una prueba depende de la naturaleza del grupo normativo, pues éste afecta en gran medida la interpretación de las puntuaciones de la prueba sin importar el tipo de normas que se obtengan. De ahí que sea importante considerar qué clase de grupos normativos podemos encontrar.

Los grupos normativos de las pruebas psicológicas muestran tal diversidad que es difícil incluirlos en categorías. Aquí presentamos un esquema de clasificación que representa puntos a lo largo de un continuo más que enfoques claramente diferenciados. En la práctica, encontraremos ejemplos de los puntos intermedios de este continuo.

Normas nacionales

Algunas pruebas aspiran a tener **normas nacionales**, es decir, normas basadas en un grupo representativo del segmento de la población nacional para la que está pensada la prueba. Este segmento podría ser todos los adultos, todos los niños de un grado específico, todos los aspirantes a entrar a la universidad o todas las personas que son legalmente invidentes. El grupo –población– meta se define, por lo general, junto con el propósito de la prueba. En el cuadro 3-4 aparece una afirmación de muestra que anuncia normas representativas de todo el país. Compara estas afirmaciones con las del cuadro 3-5, que rechazan la representatividad de cualquier grupo bien definido.

Cuadro 3-4. Afirmaciones de muestra que anuncian normas representativas de una población

Las normas del Scranton Test de Inteligencia no verbal se basan en muestras representativas elegidas cuidadosamente de niños de 3 a 16 años de edad de la población de EUA.
El University Attitude Scale tiene normas que reflejan a la población de estudiantes universitarios de la nación.

Cuadro 3-5. Afirmaciones de muestra que no anuncian normas representativas de una población

Las normas de esta prueba se basan en todas las personas que la respondieron y cuyas puntuaciones fueron procesadas por la editorial en los tres años más recientes.
Las normas del Scranton Anxiety Test se basan en 236 casos que respondieron a la prueba en el centro comunitario de consejería de Scranton.

Normas internacionales

En el contexto de los estudios internacionales de rendimiento escolar, se han elaborado normas internacionales en años recientes. Las normas se basan en niños de escuelas tomados de grupos de países que decidieron participar en estos estudios. La mayoría de las interpretaciones se basan en comparaciones de puntuaciones totales y porcentajes de estudiantes que responden correctamente a reactivos individuales. En el capítulo 11 se pueden ver ejemplos de normas internacionales.

Grupos por conveniencia para la obtención de normas

Algunas pruebas aspiran a tener normas nacionales, pero no pretenden tenerlas en realidad; en lugar de ello, tienen normas basadas en uno o varios grupos **por conveniencia**, es decir, grupos que están "convenientemente" disponibles para la aplicación de la prueba. A menudo estos grupos provienen de una sola ubicación geográfica, son relativamente homogéneos en cuanto a los antecedentes culturales y pueden tener un rango limitado de edad, nivel educativo y otras variables importantes. Algunas pruebas presentarán varias normas diferentes para distintos grupos. Por ejemplo, una prueba de autoconcepto puede presentar una norma basada en 250 estudiantes de octavo grado en una ciudad del noreste, otra norma basada en 150 personas de 15 a 18 años de edad enviados para orientación y otra norma más basada en 200 adultos que participaron en un estudio de actitudes del consumidor.

En el mejor de los casos, el usuario de la prueba espera que el manual contenga una descripción franca y detallada de las características de estos grupos de normalización *ad hoc*, pero a veces ni siquiera eso está disponible. Las normas basadas en los grupos por conveniencia deben interpretarse con sumo cuidado. El usuario debe abstenerse de suponer que tales normas pueden usarse como un buen sustituto de una norma nacional o de un subgrupo definido con precisión.

Normas del usuario «67a

Algunas pruebas emplean lo que llamamos **normas del usuario**, las cuales se basan en los grupos que en realidad respondieron la prueba, por lo general, dentro de un tiempo especificado. Conforme nuevos grupos responden la prueba, la editorial sólo agrega estos casos en su base de datos normativos. De ahí que este tipo de normas suelan encontrarse sólo en casos en que la editorial de la prueba califica todos o al menos una parte considerable de las pruebas aplicadas. Las normas de rango percentil del SAT y ACT son normas del usuario (véase capítulo 9), pues se basan en todos los estudiantes que respondieron la prueba dentro de un periodo reciente. Las normas del Major Field Test (véase capítulo 11) se basan en todos los individuos que respondieron la prueba en un periodo de tres años.

En las normas del usuario no hay un intento *a priori* de asegurarse de que el grupo sea representativo de una población bien definida. En realidad, estas normas son un tipo de

normas por conveniencia. Como señalamos en las normas por conveniencia, se espera que una detallada descripción acompañe las normas del usuario.

Normas de subgrupos

Algunas pruebas incluyen **normas de subgrupos**, que se toman del grupo total de normalización. Por ejemplo, pueden incluirse normas separadas de acuerdo con sexo, raza, nivel socioeconómico, ocupación o región geográfica.

Las normas de subgrupos pueden ser útiles sólo si hay diferencias considerables entre los subgrupos en las variables que mide la prueba. Si los subgrupos no difieren en la variable, estas normas no serán diferentes de las normas basadas en el grupo total.

Dependiendo del propósito de la aplicación de la prueba, se puede preferir emplear sólo la norma del grupo total o la del subgrupo. En muchas situaciones, usar ambas mejora la interpretación de la prueba; por ejemplo, puede ser útil saber que la puntuación de Zeke se ubica en el percentil 60 de la norma nacional, pero en el 30 de las personas que tienen la misma ocupación que él.

Normas locales

Una escuela emplea los *Metropolitan Achievement Tests*. Las puntuaciones de los estudiantes de esta escuela se informan en términos de las normas nacionales; además, la escuela prepara una distribución de las puntuaciones de sus propios estudiantes e interpreta la de cada uno en relación con las puntuaciones de los demás alumnos de la escuela. Esto se denomina **norma local**; estas normas casi siempre se expresan como percentiles.

Consideremos otro ejemplo. Una compañía usa una prueba de aptitud cuantitativa para elegir empleados de oficina. Cada año, la compañía evalúa a 200 solicitantes del trabajo. Aunque existen normas nacionales de la prueba, la compañía usa las 200 aplicaciones de la prueba para crear una norma local.

Las normas locales pueden ser de utilidad para algunos propósitos de interpretación. Una ventaja de las normas locales es que sabemos con certeza las características del grupo normativo, ya que es precisamente gente del contexto local. Desde luego, en una norma local, la persona típica estará en el promedio, lo cual puede ser engañoso.

Por ejemplo, en la situación escolar de evaluación anteriormente mencionada, el estudiante típico de cada grado estará "en la norma", lo cual no es muy informativo, ya que es una afirmación verdadera por definición. No se puede determinar con una norma local si el individuo típico está arriba o abajo del promedio en términos de algún marco de referencia externo.

La figura 3-21 muestra un ejemplo. En este caso, en que el grupo local está arriba de la norma nacional, una puntuación natural de "X" se ubica en el percentil 55 de la norma nacional, pero en el 45 de la norma local. Cuando el grupo local está muy por encima o

muy por debajo del promedio, se observará una mayor diferencia.

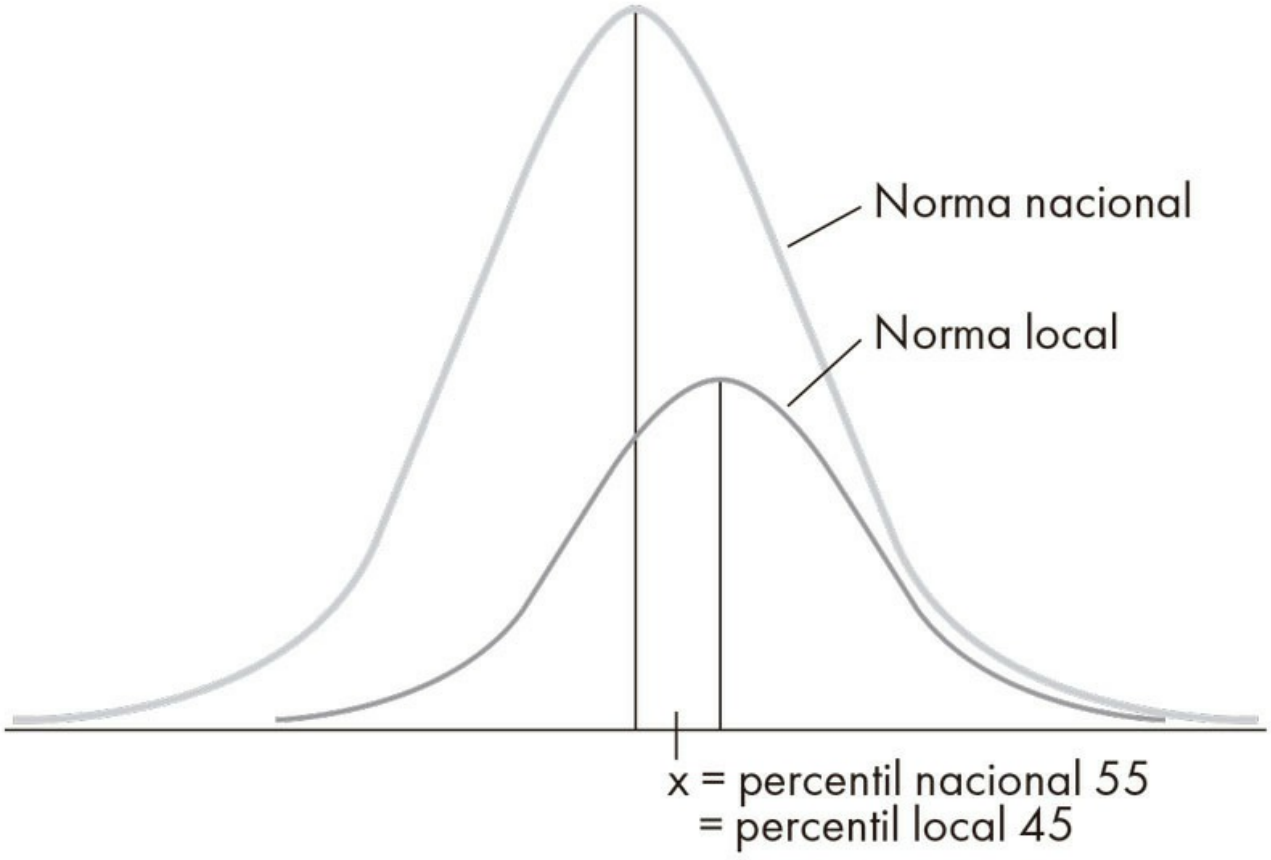


Figura 3-21. Comparación de muestra entre las normas nacional y local.

Normas institucionales

Algunas pruebas, en especial las de aprovechamiento, incluyen normas basadas en instituciones y en individuos. Las **normas institucionales** se basan en los promedios de los individuos que están dentro de las instituciones; por ejemplo, una prueba se aplica a 5000 estudiantes en 200 universidades y se obtiene el promedio de cada una de ellas y una distribución de frecuencias de dichos promedios. Con base en esta distribución, por lo general, se elabora una norma de percentiles, la cual constituye una norma institucional. También podría llamarse norma escolar, norma grupal o de alguna otra manera.

Por lo común, la distribución de puntuaciones individuales y promedios grupales tendrá aproximadamente el mismo centro, pero las puntuaciones individuales serán mucho más variadas que los promedios grupales. De ahí que una puntuación natural, excepto si está a la mitad de la distribución, estará más alejada de la norma de las instituciones que de la norma de los individuos. Las puntuaciones arriba del promedio estarán más arriba en la norma institucional que en la individual y viceversa en el caso de las puntuaciones debajo del promedio. Las diferencias pueden ser radicales. En la figura 3-22 se describe una

comparación entre normas individuales e institucionales. Éste es sólo un ejemplo; el grado real en que se traslapan ambos tipos de normas varía en distintas pruebas y grupos de normalización.

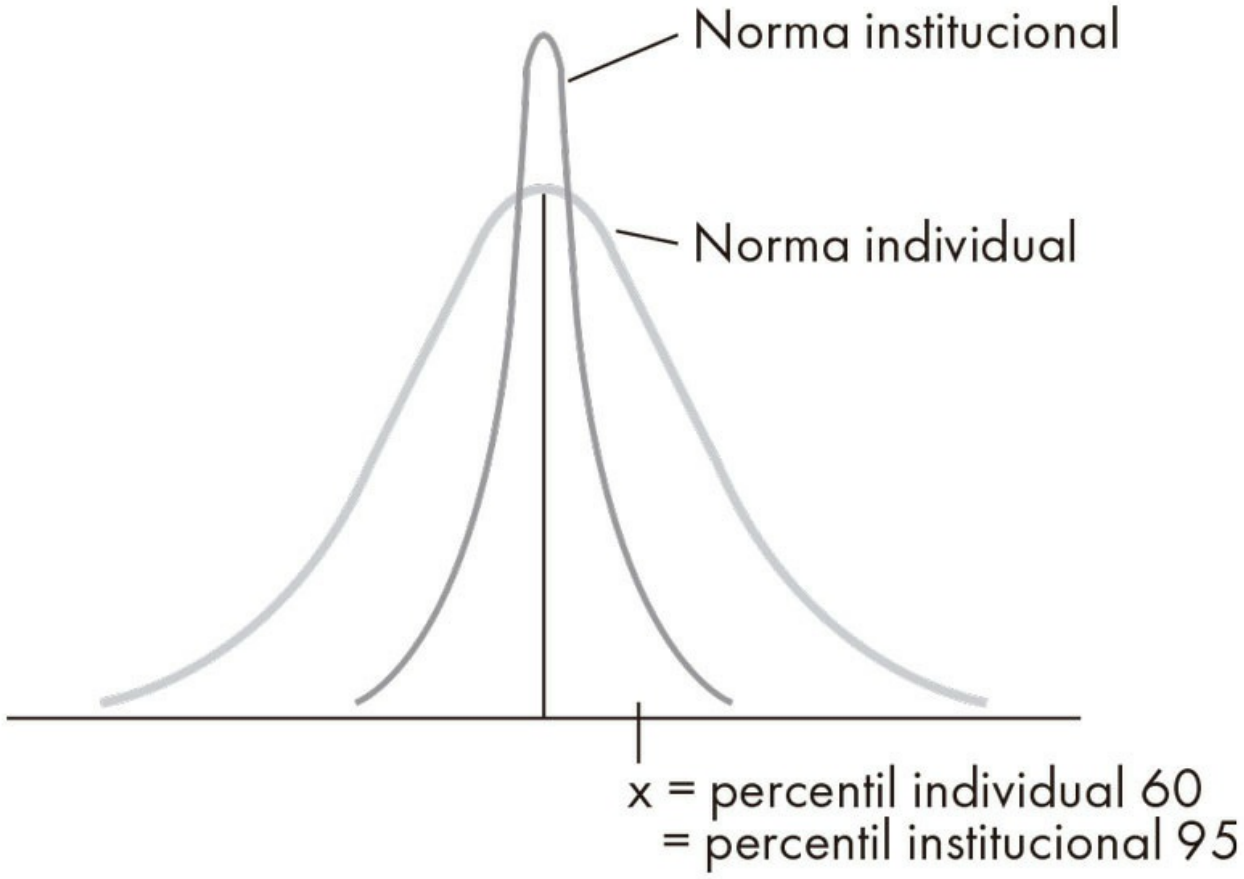


Figura 3-22. Norma individual contra norma institucional.

Puede haber mucha confusión si no distinguimos con cuidado las normas individuales de las institucionales. Consideremos, por ejemplo, la afirmación "la puntuación del suroeste se ubica en el percentil 95 de las normas nacionales". Mucha gente podría interpretar que el estudiante típico del suroeste tiene mejores puntuaciones que 95% de los estudiantes del país. Sin embargo, si el percentil 95 se basa en normas institucionales, bien podría ser que el estudiante típico del suroeste tuvo una puntuación mejor que sólo el 70% de los estudiantes del país.

Resumen de puntos clave 3-5

Principales tipos de grupos normativos
Normas nacionales
Normas internacionales
Grupos normativos por conveniencia

Normas del usuario
Normas de subgrupos
Normas locales
Normas institucionales

Interpretación referida al criterio

Una prueba de 50 reactivos de habilidad de cálculo básico se aplicó a un grupo de adultos. La prueba incluye reactivos como 7 × 9 = _____, 417 + 236 = _____ y 2596 – 1688 = _____. Un individuo tiene 30 reactivos (60%) correctos; un maestro o un padre de familia podría juzgar este desempeño como "insatisfactorio". Éste es un ejemplo de una interpretación **referida a un criterio**. Al hacer el juicio de que 60% de respuestas correctas es insatisfactorio, no hay referencia a ningún grupo normativo. Así, la interpretación referida a un criterio contrasta con la interpretación **referida a la norma**, la cual fue el tema de las primeras secciones de este capítulo.

Las pruebas mismas a veces son calificadas como referidas a un criterio o a una norma, pero esta terminología es inapropiada. No es la prueba sino el marco en que se interpreta lo que se puede calificar de estos modos. De hecho, ambos tipos de interpretación pueden usarse para la misma prueba; por ejemplo, en la prueba descrita en el párrafo anterior, podemos notar que una puntuación natural de 30 ubica al individuo en el percentil 75 de un grupo normativo que representa a la población adulta general de EUA.

La interpretación con referencia a un criterio suele ser aplicable sólo a algunos dominios de contenido bien definidos, como computación, ortografía o cualquier habilidad requerida en cierta ocupación. A menudo aplicamos este tipo de interpretación a puntuaciones de exámenes para otorgar licencias profesionales y en exámenes de competencia mínima para la graduación del bachillerato. Mientras menos definido esté el dominio, más difícil es la interpretación con referencia a un criterio. El método de evaluación también es importante cuando se aplica esta interpretación; por ejemplo, en la prueba de cálculos mencionada antes, la interpretación del desempeño puede variar de manera considerable dependiendo de si los reactivos de la prueba eran de respuesta libre o de opción múltiple, si se usó una calculadora y si había límite de tiempo. ¿Cómo podría cambiar la interpretación de "60% correcto" si no había límite de tiempo o cada reactivo tenía que responderse en 10 segundos? La interpretación con referencia a un criterio a menudo suena como una idea simple, pero puede volverse problemática cuando se examina con mayor detenimiento.

La interpretación con referencia a un criterio a menudo se aplica a los exámenes de salón de clases. El conocido esquema de calificación "90% es una A, 80-89% es una B,... menos de 60% es no aprobado" es un ejemplo de esta interpretación. El dominio de contenido puede definirse como "lo que el maestro revisó en clase" o "todo lo que aparece en los capítulos 1 al 4 del libro de texto". El criterio es que los estudiantes debieron aprender todo o casi todo este material; luego, se aplica el conocido esquema de

calificación. Por supuesto, las calificaciones pueden estar "curveadas", de modo que la calificación promedio es "C" y así sucesivamente. Tal "curvatura" es, en realidad, un tipo de interpretación con referencia a una norma, en particular, la aplicación de una norma local.

Los **niveles de competencia** establecidos para las evaluaciones estatales del aprovechamiento ofrecen otro buen ejemplo de la interpretación con referencia a un criterio. Estos niveles se denominan a menudo *estándares de desempeño*. Un sistema común emplea cuatro categorías: Avanzado, Competente, Básico e Inferior al básico. Las comisiones, por lo general integradas por maestros, administradores escolares y representantes del público, examinan el contenido de las pruebas y hacen juicios sobre qué puntuación natural en la prueba justifica llamar "competente" o "avanzado" a un estudiante. Así, se necesitan tres puntos de corte para separar las cuatro categorías mencionadas arriba. Se han desarrollado numerosos procedimientos específicos para crear tales puntos de corte, pero todos fallan frente a un juicio con referencia a un criterio sobre el desempeño. En Cizek (2001b) se puede encontrar la descripción de una gran variedad de métodos, y en Cizek, Bunch y Koons (2004) se encuentra una introducción práctica a los métodos. Usar estas categorías de competencia no excluye usar también las normas; por ejemplo, podríamos encontrar que el punto de corte que separa las categorías básico y competente corresponde al percentil 45 de la prueba.

Grupo de estandarización: determinar su utilidad

Aplicar una prueba a los individuos del grupo normativo en lo que se domina programa de obtención de normas o de **estandarización**. Éste es, por lo común, uno de los últimos pasos en la elaboración de pruebas. En el capítulo 6 se describe con detalle el proceso entero de elaboración de pruebas. Aquí tratamos la cuestión de cómo determinar si un programa de estandarización es bueno. El uso competente de las pruebas psicológicas requiere no sólo de conocimiento sobre los tipos de normas –percentiles, puntuaciones estándar– sino también de la capacidad para juzgar la utilidad de las normas. Por ejemplo, pensemos, utilidad se refiere al grado en que las normas proporcionan un marco significativo para interpretar la prueba. En la mayoría de casos, esto significa tener una norma que es 1) estable y 2) representativa de una población bien definida. De ahí que existan dos temas: estabilidad y representatividad.

La **estabilidad** de una norma se determina en gran medida por el tamaño del grupo normativo, es decir, el número de casos en el programa de estandarización. Esto rara vez es un problema, pues no se necesitan muchos casos para alcanzar la estabilidad estadística. Algunos cientos de casos producirán suficiente estabilidad para la mayoría de usos prácticos de las normas. Consideremos la norma de un sistema de puntuación estándar con $M = 100$ y $DE = 15$; con $N = 300$, el error estándar de la media es menor de un punto y el intervalo de confianza de 95% es $+/-1.7$ puntos; ésa es una buena estabilidad. En la práctica, las normas de muchas pruebas se basan en miles de casos.

Al considerar el número de casos de un grupo de normalización, necesitamos

determinar el tamaño de dicho grupo en que se basa una norma específica. El número total de casos agregados en los distintos grupos normativos no es el número decisivo. Por ejemplo, una prueba puede jactarse de que sus normas se basan en casi 1000 casos; sin embargo, supongamos que las normas reales aparecen por separado para cada género y cada uno de 10 grados. Así, en realidad hay 20 grupos normativos con 50 casos cada uno; por lo tanto, aquí el número importante para determinar la estabilidad es 50 y no 1000.

Como lo señalamos, la estabilidad rara vez es un problema y, en cualquier caso, se puede determinar con facilidad. Sin embargo, la estabilidad no garantiza representatividad; la diferencia entre estos dos conceptos es lo más importante que se debe aprender en relación con las normas. Es posible tener grupos normativos muy grandes que producen normas sumamente estables, pero que no son representativos de la población meta de la prueba.

¿Cómo determinaremos la *representatividad* de un grupo de normalización? La respuesta a esta pregunta depende de lo que el autor de la prueba afirme sobre las normas. Hay dos posibilidades; primero, el autor puede afirmar que las normas son representativas de una población particular. Por ejemplo, puede afirmar que el grupo de estadarización es representativo de todos los adultos de EUA de 20 a 80 años de edad, o de todos los estudiantes de sexto grado, o de todas las mujeres universitarias. Segundo, el autor puede no afirmar que las normas son representativas de alguna población particular, sino que sólo presentar la muestra normativa como una norma de conveniencia o de un grupo de usuarios.

Ahora consideremos cómo juzgar la representatividad del grupo normativo en el primer caso, en que se afirma que dicho grupo es representativo de la población meta. En cierto modo, determinar la cualidad de un grupo de estandarización que se pretende sea representativo de una población es tema de la teoría del muestreo. Sin duda, el lector conoce las técnicas de muestreo aleatorio que se estudian en los cursos de estadística básica; sin embargo, este muestro se usa rara vez para obtener las normas de una prueba. Por lo común, se recurre a alguna forma del muestreo estratificado por grupos. La cuestión del muestreo siempre es complicada debido a que la participación en el programa de estandarización suele ser voluntaria, por lo que, en la práctica, en vez de concentrar nuestra atención en las nociones de la teoría del muestreo, tenemos que enfocarnos en la evidencia relacionada con la correspondencia entre el grupo de estandarización y la población meta tomando en cuenta características importantes. En este contexto, dichas características se refieren a la variable que se quiere medir; con frecuencia, se emplean diversas características demográficas para mostrar esta correspondencia. Por lo general, las características empleadas (véase cuadro 3-6) son edad, género, raza/etnia, estatus socioeconómico y región geográfica. También se emplea el desempeño en otras pruebas que tienen normas bien documentadas. Debe demostrarse que el grupo de estandarización corresponde a la población meta suficientemente bien en tales características.

Cuadro 3-6. Tipos de información útil para juzgar la utilidad de un grupo de normalización

Edad	Grupo racial/étnico
Género	Estatus socioeconómico
Nivel de capacidad	Región geográfica
Nivel educativo	Tamaño de la ciudad

A menudo, cuando no hay correspondencia entre el grupo de estandarización y la población meta, se les agregará peso a los casos de dicho grupo para mejorar la correspondencia. Supongamos, por ejemplo, que la población tiene 50% de hombres y 50% de mujeres, mientras que el grupo de estandarización tiene 40% de hombres y 60% de mujeres. Asignar un peso de 1.5 a cada hombre (o de .67 a cada mujer) producirá un grupo compensado de estandarización con la proporción 50-50 requerida en relación con el género.

La figura 3-23 muestra ejemplos de características de un grupo de normalización que se comparan con las estadísticas nacionales. El grupo de estandarización del ejemplo A tiene una muy buena correspondencia con las estadísticas nacionales, al menos en las dos características que se muestran, mientras que la correspondencia del ejemplo B es pobre tanto en región geográfica como en nivel educativo. El cuadro nos da idea de qué información puede esperar el usuario de las pruebas que afirman tener normas representativas a nivel nacional. Desde luego, el manual de la prueba contendrá información adicional sobre el proceso de estandarización.

> Los informes de los estudios de obtención de normas deben incluir especificaciones precisas de la población de la que se obtiene la muestra, procedimientos de muestreo, índices de participación, cualquier compensación de la muestra, fechas de aplicación y estadísticas descriptivas. La documentación técnica debe indicar la precisión de las normas mismas.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Característica	Población*	Ejemplo A	Ejemplo B
Región geográfica	%	%	%
Noreste	17.9.	16.2	39.2
Medio oeste	21.7	23.4	32.1
Sur	37.1	35.0	16.4
Oeste	23.3	25.4	12.3
Nivel educativo	%	%	%
Bachillerato inconcluso	12.9	12.1	4.5
Bachillerato terminado	31.2	29.6	9.8
Estudios universitarios inconclusos	16.8	15.4	12.6
Pasante	9.1	8.0	10.4
Licenciado	19.4	22.3	30.5
Posgrado	10.5	12.6	32.2

* Datos de acuerdo con el Censo 2010 de EUA.

Figura 3-23. Ejemplos de información demográfica para dos programas de estandarización.

Cuando se afirma que un grupo de estandarización es representativo de una población particular, es responsabilidad del autor de la prueba proporcionar información suficiente que justifique tal afirmación. El usuario necesita ser especialmente precavido frente a este tipo de afirmaciones cuando la información sobre la correspondencia entre el grupo de estandarización y la población es mínima en características importantes. Por ejemplo, sólo mostrar que un grupo de estandarización para una prueba de inteligencia corresponde a la población nacional en términos de edad y género es insuficiente. ¿Qué hay del estatus socioeconómico y el nivel educativo? Debemos ser precavidos frente a

declaraciones como: “Las normas se basan en una muestra grande de 1250 estudiantes de tres universidades urbanas en el sureste de EUA”. Sí, 1250 casos conforman una muestra grande, pero eso no dice prácticamente nada acerca del grupo de estandarización, como su nivel de capacidad, nivel socioeconómico o composición racial.

Incluso cuando un grupo de estandarización muestra una buena correspondencia con la población meta en características importantes, existen dos problemas que infestan el proceso de obtención de normas. El primero es efecto de la no participación, es decir, surge del hecho de que la participación en un programa de estandarización casi siempre es voluntaria, sea del individuo o de la organización a la que éste pertenece. ¿Qué clase de individuos u organizaciones no aceptaron participar? ¿Cuáles son sus características? ¿Qué efecto puede tener su no participación en las normas? A menudo no tenemos muy buenas respuestas a estas preguntas.

Segundo, los programas de estandarización son programas de investigación más que usos ordinarios de las pruebas, y los participantes, por lo general, lo saben. En estas circunstancias es difícil asegurar que los niveles de motivación de los participantes sean los mismos que cuando se trata de una aplicación ordinaria de la prueba. Al igual que con la no participación, a menudo no conocemos el efecto del nivel de motivación en las normas. Lo mejor que podemos esperar es una discusión franca de estos problemas en el manual de la prueba.

Ahora consideremos nuestro segundo caso: cuando no se afirma que el grupo normativo es representativo de una población. ¿Qué criterios emplearemos para juzgar la calidad de las normas? Básicamente, invertimos el proceso anterior. En el primer caso, teníamos una población meta y tratábamos de demostrar que el grupo de estandarización era representativo de ella; la demostración dependía de la información sobre características importantes. En este segundo caso, esperamos tener información adecuada sobre el grupo normativo de modo que podamos proyectar qué población podría reflejarse en dicho grupo. Por ejemplo, si tenemos normas de usuario de una prueba de aprovechamiento empleada en universidades, desearíamos información sobre las universidades, como tamaño, índices de aceptación y retención, datos curriculares, género o raza/etnia. Con base en esta información descriptiva, podríamos inferir que las normas corresponden a una academia de artes pequeña, selectiva y liberal, pero no a una universidad grande, urbana, de admisión abierta. Si tenemos normas por conveniencia de una prueba de autoconcepto para estudiantes de bachillerato, desearíamos saber sobre el nivel de capacidad, género, raza/etnia y estatus socioeconómico de los participantes. Con base en esta información, podríamos concluir que las normas probablemente corresponden a escuelas de nivel socioeconómico bajo, muy urbanizadas, pero, sin duda, no se trata de un grupo representativo a nivel nacional.

Aquí hay un punto importante (y contrario al sentido común) relacionado con la composición de los grupos normativos.. El desglose del grupo en distintas características, como edad o género, es importante sólo en la medida en que los subgrupos difieren en la característica que mide la prueba. Por ejemplo, en el caso de una prueba de autoconcepto, si hombres y mujeres no difieren en esta variable, la combinación de

ambos géneros en un grupo normativo es irrelevante. Se podrían hacer las normas de la prueba sólo con hombres (o con mujeres) sin afectar la calidad de las normas. Sería un desastre político, pero, desde el punto de vista de la psicometría, sería por completo aceptable.

Resumen

1. La puntuación natural es, por lo general, el resultado más inmediato de una prueba. Por lo común, esta puntuación es el número de respuestas correctas en las pruebas cognitivas y el número de respuestas en cierta dirección en pruebas no cognitivas.
2. Las pruebas que se califican de acuerdo con los métodos TRR y que producen una puntuación theta (θ) toman en cuenta el nivel de dificultad de los reactivos y, a veces, los patrones de respuesta.
3. La distribución de las puntuaciones naturales en un grupo de estandarización constituye la base para convertirlas en puntuaciones normativas, las cuales ayudan a dar significado a las puntuaciones naturales.
4. Para entender las normas de las pruebas, se necesita tener conocimiento sobre los siguientes temas de estadística básica: distribuciones de frecuencia, formas de las distribuciones, medidas de tendencia central y de variabilidad, y puntuaciones z dentro de la distribución normal.
5. El rango percentil indica el porcentaje de individuos del grupo normativo que obtuvieron puntuaciones debajo de una puntuación natural determinada.
6. Las puntuaciones estándar son normas que convierten la distribución de puntuaciones naturales en una distribución con una media y una desviación estándar nuevas y prácticas. Existen varios sistemas de puntuación estándar que se usan mucho.
7. Las normas de desarrollo expresan el desempeño en términos de la puntuación que es típica para una edad o grado. Las normas de desarrollo más comunes son la edad mental y los equivalentes de grado.
8. Cada tipo de norma tiene sus propias ventajas y desventajas.
9. La calidad de la norma depende de las características del grupo normativo. Lo más importante es el grado en que este grupo es representativo de una población bien definida.
10. Los tipos comunes de grupos normativos son: nacional, por conveniencia, del usuario y local. Sólo unas pocas pruebas tienen normas internacionales. También hay normas de subgrupos e institucionales en el caso de algunas pruebas.
11. Algunas pruebas pueden interpretarse con referencia a un criterio en lugar de con referencia a una norma. La interpretación con referencia a un criterio consiste en un juicio relativamente directo sobre la calidad del desempeño en la prueba sin referencia a ningún grupo normativo. En el caso de algunas pruebas, se pueden utilizar ambos tipos de interpretación.
12. Es importante ser capaz de juzgar la utilidad de un grupo normativo. La información sobre las características de este grupo es crucial para hacer tales juicios. Al hacer esto, a menudo es útil comparar el grupo de estandarización con una población en términos de características como edad, género, grupo étnico/racial, región geográfica y características socioeconómicas como nivel educativo e ingreso familiar.

Palabras clave

asimetría
cero absoluto
CI de desviación
CI de razón
con referencia a un criterio
constructo
curtosis
curva normal
desviación estándar
distribución de frecuencias
edad cronológica
edad mental
efecto Barnum
equivalente de curva normal
equivalente de edad
equivalente de grado
escala de razón
escala intervalar
escala nominal
escala ordinal
estadística descriptiva
estadística inferencial
estandarización
estanina o eneatipo
grupo de normalización
grupo por conveniencia
histograma de frecuencias
informe interpretativo
media
mediana
moda
norma de desarrollo
norma del usuario
norma de subgrupos
norma institucional
norma local
norma nacional
percentil
polígono de frecuencias

pruebas referidas a la norma
pruebas referidas al criterio
puntuación de porcentaje correcto puntuación escalar
puntuación estándar
puntuación estándar normalizada
puntuación natural
puntuación normativa
puntuación T
puntuación z
rango
rango intercuartil
rango percentil
tendencia central
theta
transformación lineal
transformación no lineal
variabilidad
variable
varianza

Ejercicios

1. Con los datos de la figura 3-1, crea una distribución de frecuencias con intervalos de 5 puntos empezando con 65-69.
2. Calcula la media, mediana y desviación estándar de estas puntuaciones:5, 3, 6, 8, 8.
M = _____
Mdn = _____
DE = _____
3. Si es factible, registra la estatura de todos tus compañeros de clase y crea una distribución de frecuencias. ¿Qué forma tiene la distribución comparada con los modelos que aparecen en la figura 3-6? Haz el mismo ejercicio, pero ahora con el pulso.
4. Usando la figura 3-10a, haz *estimaciones* de los valores faltantes.

puntuación z = +1.0	Percentil = _____	ECN = _____	CI Wechsler = _____
Percentil = 75	puntuación z = _____	Otis-Lennon = _____	estanina = _____
puntuación T = 30	Percentil = _____	estanina = _____	puntuación z = _____

5. Usando el cuadro 3-1, llena los valores exactos de los mismos casos (*Nota*: en el caso del CI Wechsler, *DE* = 15, y en el del CI Otis-Lennon, *DE* = 16).

puntuación z = +1.0	Percentil = _____	ECN = _____	CI Wechsler = _____
Percentil = 75	puntuación z = _____	Otis-Lennon = _____	estanina = _____
puntuación T = 30	Percentil = _____	estanina = _____	puntuación z = _____

6. Consulta la figura 3-12. Convierte la puntuación natural de 32 al sistema de puntuaciones estándar.
7. Con base en la figura 3-16, ¿cuál es la edad mental estimada de una persona cuya puntuación natural es 35?
8. Con base en el cuadro 3-3, ¿qué rango percentil y qué puntuación estándar (CID) corresponden a una puntuación natural de 59?
9. Inventa un conjunto de categorías que puedas usar en un informe interpretativo para cada cuartil, es decir, percentiles 1-25, 26-50, 51-75 y 76-99. Las categorías pueden consistir en más de una palabra. Para evitar el uso repetido de una categoría, propón dos categorías equivalentes para cada cuartil. El programa de cómputo escogerá, entonces, uno de manera aleatoria cada vez que se aluda al cuartil.

Cuartil	**Categoría verbal A**	**Categoría verbal B**
1-25	_____	_____
26-50	_____	_____

51-75	_____	_____
76-99	_____	_____

10. Introduce los datos del apéndice D1: GPA en una hoja de cálculo de un paquete estadístico como SPSS, SAS o Excel. Corre el programa para obtener las medias y desviaciones estándar de cada variable. Por cada dos variables, obtén distribuciones de frecuencia e histogramas. ¿Cómo describirías las formas de las distribuciones?

11. Entra al sitio web de estas editoriales para examinar los informes interpretativos de pruebas muy usadas. Escribe descripciones breves de cómo piensas que el programa de cómputo generó tales informes.

- http://psychcorp.pearsonassessments.com/pai/ca/research/publications/samplerpts/
- http://www4.parinc.com/WebUploads/samplerpts/PAI32.pdf
- http://portal.wpspublish.com/portal/page_pageid=53,124601&_dad=portal&_sche

O también puedes escribir "interpretive report" o "interpretive report for [agregas el nombre o el acrónimo de la prueba]" en cualquier buscador de internet para acceder a una gran cantidad de muestras de informes interpretativos.

Notas

[1] En el capítulo 1 describimos el importante papel de *Standards for Educational and Psychological Testing*, publicado en conjunto por *American Educational Research Association, American Psychological Association* y *National Council on Measurement in Education*. Los capítulos 3 al 6 contienen extractos breves de este documento, abreviado a menudo en la literatura como "Standards", para ilustrar ciertos puntos que se tocan en el texto.

[2] El símbolo σ (sigma) es la letra griega (minúscula) que corresponde a la S. En estadística, usamos las letras griegas para designar medidas de la población entera y las letras latinas para designar las medidas de las muestras. Así, DE representa la desviación estándar de una muestra, mientras que σ, la desviación estándar de la población. La distinción no se respeta siempre en la literatura de las pruebas psicológicas.

[3] Esta ecuación genera una curva normal:

$$Y = (N/(\sigma\sqrt{2\pi})) \left(e^{-(X-\mu)^2/2\sigma^2}\right)$$

[4] Si el lector desea profundizar en el tema de los percentiles, consúltese el cap. 4 del libro "Evaluación Psicológica. Historia, fundamentos teórico conceptuales y psicometría", 2a edición, de Laura Edna Aragón Borja, Ed. El Manual Moderno.

CAPÍTULO 4
Confiabilidad

Objetivos

1. Definir confiabilidad tal como se usa el término en las pruebas psicológicas.

2. Refrescar tu conocimiento de conceptos estadísticos básicos relacionados con la correlación y predicción, incluyendo factores que afectan la magnitud de las correlaciones.

3. Distinguir entre confiabilidad y validez, entre distintos usos cotidianos del término confiabilidad, entre cambios reales y fluctuaciones temporales, y entre errores constantes y errores no sistemáticos.

4. Identificar las principales fuentes de falta de confiabilidad de las puntuaciones de las pruebas.

5. Describir los componentes de la teoría de la puntuación verdadera.

6. Para cada uno de estos métodos de confiabilidad, decir cómo se lleva a cabo un estudio y qué factores que afectan la confiabilidad atacan: test-retest, interjueces, formas paralelas, consistencia interna.

7. Definir y calcular el error estándar de medición y los intervalos de confianza.

8. Distinguir el error estándar de medición del error estándar de la media y del error estándar de estimación.

9. Definir qué significa precisión de la medición en la TRR.

10. Describir qué teoría de la generalizabilidad se describe.

11. Determinar cómo los factores que afectan el coeficiente de correlación influyen en los datos de confiabilidad.

12. Ofrecer puntos de referencia de los niveles aceptables de confiabilidad.

Introducción

Juan hace la prueba de admisión a la universidad el sábado 2 de octubre, después de una dura semana en la escuela, coronada por un partido de futbol la noche del viernes. También Raúl hace la prueba el 2 de octubre; se siente filoso como cuchillo y dispuesto a comerse el mundo. ¿Juan y Raúl obtendrán puntuaciones considerablemente distintas si presentan el examen el sábado 9 de octubre, cuando sus circunstancias personales sean diferentes?

La clase de química de Tomás incluye a 700 de nuevo ingreso. Para evitar que hagan trampa durante el examen, el profesor toma 100 problemas y los divide en cuatro para tener las formas A, B, C y D del examen, con 25 problemas cada una. Reparte al azar las formas. ¿La calificación de Tomás será muy distinta si le toca la forma A o la forma B?

¿Qué tanto fluctúan las puntuaciones en una prueba de personalidad de un día a otro? ¿Qué tan parecidas son las calificaciones de un ensayo dependiendo de quién lo califique? Cuando dos clínicos emplean una forma para valorar la gravedad de un desajuste psicológico, ¿es probable que concuerden en sus valoraciones?

Todas estas preguntas se relacionan con el tema de la confiabilidad. Este capítulo considera cómo responder a esta clase de preguntas. Antes de empezar nuestro tratamiento formal de la confiabilidad, debemos hacer cuatro distinciones importantes.

Cuatro distinciones importantes

Primero, debemos distinguir entre *confiabilidad* y *validez* de las medidas. La validez se tratará de manera más completa en el siguiente capítulo, pero la definiremos brevemente aquí para contrastarla con la confiabilidad. La validez se ocupa de lo que mide una prueba, es decir, si mide lo que pretende medir; en cambio, la **confiabilidad** se ocupa sólo de la consistencia de la medida, sin importar qué es, con precisión, lo que se está midiendo. Una medida puede ser confiable sin ser válida; por ejemplo, la prueba de química a la que nos referíamos antes puede ser muy confiable, pero puede ser más una medida de habilidad matemática que de conocimiento sobre química. Puede haber un excelente consenso entre los clínicos que valoran un desajuste, pero la forma de valoración puede ser más una medida de habilidad verbal pobre que una de desajuste. Aunque una prueba puede ser confiable sin ser válida, no puede ser válida a menos que sea confiable. En este capítulo, nos ocuparemos sólo del tema de la confiabilidad.

Segundo, debemos estar conscientes de las diferencias entre los *usos cotidianos de la palabra* confiabilidad y su uso técnico en el campo de las pruebas psicológicas. En lenguaje cotidiano, la palabra *confiabilidad* tiene varios significados relacionados entre sí. Una máquina confiable inicia y funciona de manera continua cuando oprimimos el botón ON. Un empleado confiable llega con puntualidad y casi nunca falta al trabajo. Una "fuente por lo general confiable" proporciona información exacta en lugar de rumores. Un vendedor de autos confiable ha estado en el negocio por años, se espera que continúe ahí y dé un buen servicio a los clientes.

Todos estos significados cotidianos son pertinentes para el concepto de confiabilidad en el campo de las pruebas psicológicas. Sin embargo, la confiabilidad de las pruebas tiene un significado más técnico y cuantitativo. Los mejores sinónimos en español para el término técnico de confiabilidad son consistencia, replicabilidad y fiabilidad. Una prueba confiable, en sentido psicométrico, produce de manera *consistente* la misma puntuación o una similar para un individuo. La puntuación puede *replicarse* al menos dentro de cierto margen de error. Nos podemos *fiar* de que una prueba confiable producirá la misma puntuación para un individuo. Este capítulo se ocupa del significado técnico, psicométrico, de la palabra *confiabilidad*.

Tercero, debe hacerse la distinción entre un **cambio real** en el rasgo que se mide y fluctuaciones en las puntuaciones que se pueden atribuir a cambios pasajeros en las circunstancias personales, la "suerte" de cada quien mientras hace la prueba o diferencias debidas a quien se encarga de calificarla. Los cambios reales no son fuente de falta de confiabilidad, mientras que los otros factores sí lo son, a menos que estemos tratando de medir cambios en el estado de ánimo o emocional. No hay una clara demarcación entre los cambios temporales a corto plazo y los cambios reales a largo plazo, pero la distinción es importante en términos conceptuales.

Cuarto, necesitamos distinguir entre errores sistemáticos o constantes y **errores no sistemáticos** en nuestras mediciones. Un **error constante** lleva a que la puntuación de

una persona sea sistemáticamente alta o baja, sin importar la constancia en el rasgo que se está midiendo en la condición que la persona esté presentando. Por ejemplo, consideremos el nivel de inteligencia de un niño cuya lengua materna es el español, pero a quien se aplica la prueba en inglés. Es probable que el nivel de inteligencia del niño se subestime, pero esta subestimación será relativamente constante si el niño es evaluado el martes o el miércoles. O consideremos a Jessica, que es buena para responder pruebas, pues sabe detectar pistas para encontrar la respuesta correcta incluso cuando no sabe mucho del tema en cuestión. Jessica tiende a obtener puntuaciones superiores a lo que su conocimiento le permitiría, y esto sucede sin importar cuándo hace las pruebas. La confiabilidad no explica estos errores constantes, pues sólo trata con errores no sistemáticos. Podemos notar que lo que llamamos errores "constantes" no son en realidad constantes, sino que son tendencias que modifican las puntuaciones en cierta dirección.

Resumen de puntos clave 4-1

Cuatro distinciones importantes de la confiabilidad

1. **1.** Confiabilidad frente a validez
2. **2.** Usos cotidianos en oposición a definición técnica
3. **3.** Cambio real frente a cambio temporal
4. **4.** Errores sistemáticos o constantes contra errores no sistemáticos

Revisión de estadística: Parte 2 – Correlación y predicción

Los coeficientes de correlación y sus derivaciones –errores estándar y fórmulas de predicción– son elementos cruciales en nuestro estudio de la confiabilidad y la validez, temas de éste y el siguiente capítulo. De ahí que será útil hacer una rápida revisión de conceptos y procedimientos clave relacionados con estos métodos estadísticos. Al igual que en la revisión de estadística en el capítulo anterior, suponemos que el lector ha tenido una introducción completa a este material, pero puede necesitar refrescar su conocimiento para activar sus viejos recuerdos.

Distribución bivariada y coeficientes de correlación

La relación entre dos variables puede representarse por medio de una **distribución bivariada**, también conocida como **dispersograma**. La figura 4-1 presenta varias de estas distribuciones; en cada caso, la variable *X* está en el eje horizontal y la variable *Y* en el eje vertical. Cada punto (•) de una distribución corresponde a las coordenadas (*X*, *Y*) de un único caso; por ejemplo, si *X* es una puntuación de la forma X de una prueba y *Y* es una puntuación de la forma Y de una prueba, entonces las coordenadas (*X*, *Y*) representan las puntuaciones de un individuo en las formas X y Y.

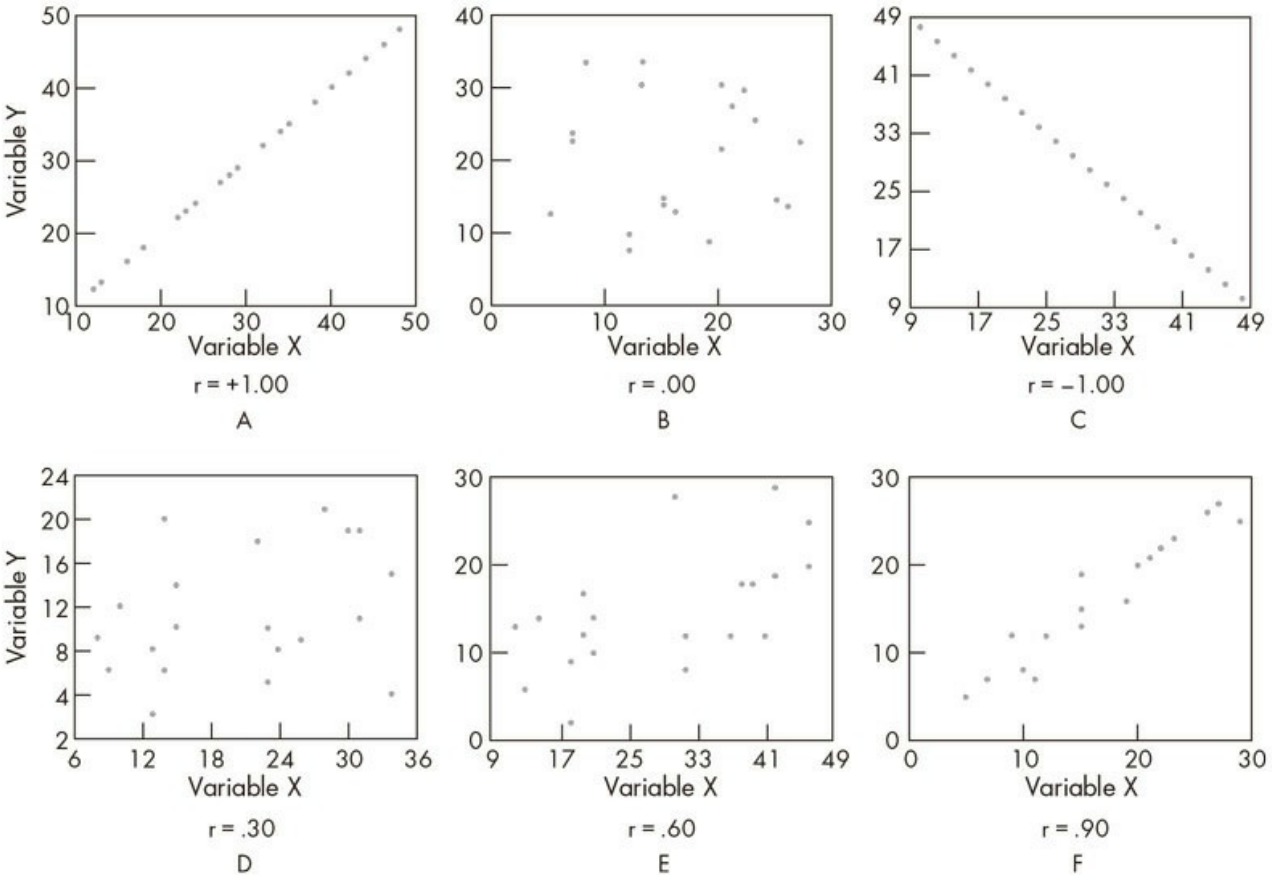


Figura 4-1. Ejemplos de distribuciones bivariadas y sus correspondientes *r*.

El **coeficiente de correlación** *r* de Pearson brinda un resumen numérico de la relación representada en una distribución bivariada. Al presentar las fórmulas de *r*, solemos distinguir entre una definición básica de *r* y un cómputo o fórmula de puntuaciones naturales. Las siguientes suelen emplearse como definiciones y versiones de cómputo de las fórmulas. Sin duda, las viste en tu curso de estadística básica.

Definición

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N S_X S_Y}$$

Fórmula 4-1

Fórmula de cómputo

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}}$$

Fórmula 4-2

El valor de *r* puede variar de −1.00 a +1.00. Una *r* de +1.00 representa una relación lineal positiva perfecta entre dos variables, como se muestra en la gráfica A de la figura 4-1. Una *r* de −1.00 representa una relación lineal negativa perfecta entre dos variables, como se muestra en la gráfica C. En cambio, una *r* de .00 representa la ausencia de relación entre las dos variables, como se muestra en la gráfica B. En la mayor parte del trabajo práctico en el campo de las pruebas psicológicas, encontramos *r* que están lejos de ser perfectas. La figura 4-1 muestra distribuciones bivariadas de distintos valores intermedios de *r*: .30, .60 y .90. La mayor parte de las correlaciones que encontramos en las pruebas psicológicas, así como en otras obras del área de las ciencias sociales y de la conducta, son correlaciones de Pearson.

Las fórmulas de *r* que presentamos antes son de esta correlación. Sin embargo, existen otros tipos de coeficientes de correlación, algunos de los cuales son variantes del de Pearson y se pueden aplicar cuando la naturaleza de la escala permite la simplificación computacional de la fórmula. Por ejemplo, cuando una variable es dicotómica y las únicas puntuaciones posibles son 0 y 1, entonces $[\Sigma]\ X / N = p$, el porcentaje de casos con puntuación de 1. Esto permite simplificar la fórmula de cómputo de *r*. Otros tipos de correlación no son simples variaciones del coeficiente de Pearson, sino que se han obtenido por otras vías. Sin embargo, todos los coeficientes de correlación pueden interpretarse de la misma manera que el de Pearson. Hay unas pocas excepciones a esta generalización, pero las excepciones rara vez tienen implicaciones prácticas importantes. El cuadro 4-1 enumera los distintos tipos de coeficientes de correlación bivariada. A menos que se especifique otra cosa, a lo largo de este libro suponemos que una correlación es la de Pearson. En el capítulo 5, examinaremos la correlación multivariada, en particular, la correlación múltiple y la correlación parcial.

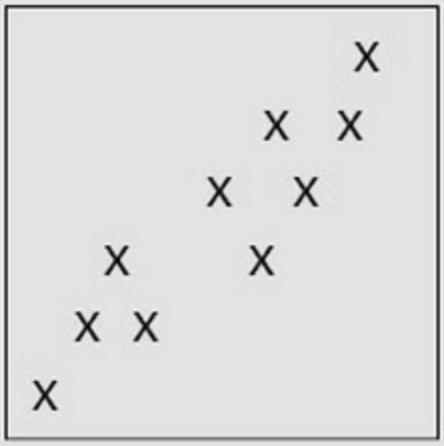
Cuadro 4-1. Ejemplos de los tipos de coeficientes de correlación distintos del de Pearson	
Biserial *r* (r_{bis})	Biserial por puntos *r*(r_{pbis})
Tetracórica *r* (r_{tet})	Coeficiente phi (Θ)
Coeficiente de contingencia (*C*)	Correlación por rangos ordenados de Spearman (R)
Correlación intraclase (*CIC*)	Eta (η)
Kappa (κ)	Tau de Kendall (τ)

Regresión lineal

Una vez establecida la correlación r entre dos variables X y Y, podemos usarla para predecir el valor de Y si conocemos el de X (o viceversa).

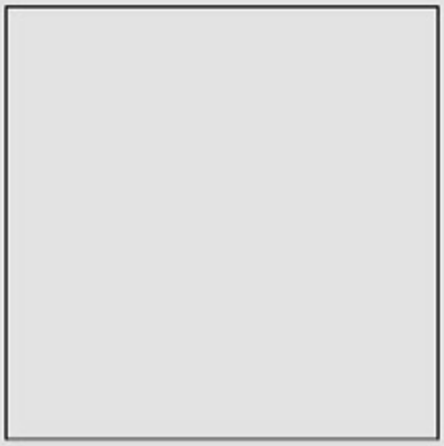
¡Inténtalo!

Calcula las *r* de las siguientes distribuciones bivariadas.



$r =$ _____

En este cuadro, llena con marcas (*) para representar una distribución bivariada correspondiente a $r = .40$.



Compara tus cálculos con los de otros estudiantes.

Supongamos que tenemos una correlación entre una prueba de admisión a la universidad (*X*) y el grade point average (GPA [promedio de las calificaciones de un grado]) (*Y*). Ahora tenemos la puntuación *X* de un estudiante y deseamos predecir su *Y'*

del GPA (usamos ’ para indicar que se trata de una *Y* predicha y no de una *Y* conocida). La forma general de la ecuación de predicción es:

$$Y' = bX + a$$

donde *b* es la pendiente de la **línea de regresión**[1] y *a* es la intersección con el el eje *y* u ordenada al origen. Ésta es la línea más adecuada de acuerdo con el criterio de los mínimos cuadrados: minimiza la cantidad

$$\sum (Y - Y')^2$$

Una fórmula de cómputo conveniente, y equivalente en términos algebraicos a la última fórmula, es:

$$Y' = r_{xy}\left(\frac{DE_y}{DE_x}\right)(X - \overline{X}) + \overline{Y}$$

Fórmula 4-3

r_{xy} = correlación entre *X* y *Y*
DE_x = desviación estándar de *X*
DE_y = desviación estándar de *Y*
X = puntuación en *X* de una persona
$\overline{X}$ = media de las puntuaciones *X*
$\overline{Y}$ = media de las puntuaciones *Y*

La figura 4-2 presenta un ejemplo de una línea de regresión. Cada punto de la figura representa las coordenadas *X* y *Y* de una persona. Usamos esta línea para predecir los valores de *Y* a partir de los de *X*; por ejemplo, la línea punteada en el cuadrante inferior de la figura. En el caso de una persona con una puntuación *X* de 9, podemos predecir una puntuación *Y* de 29.

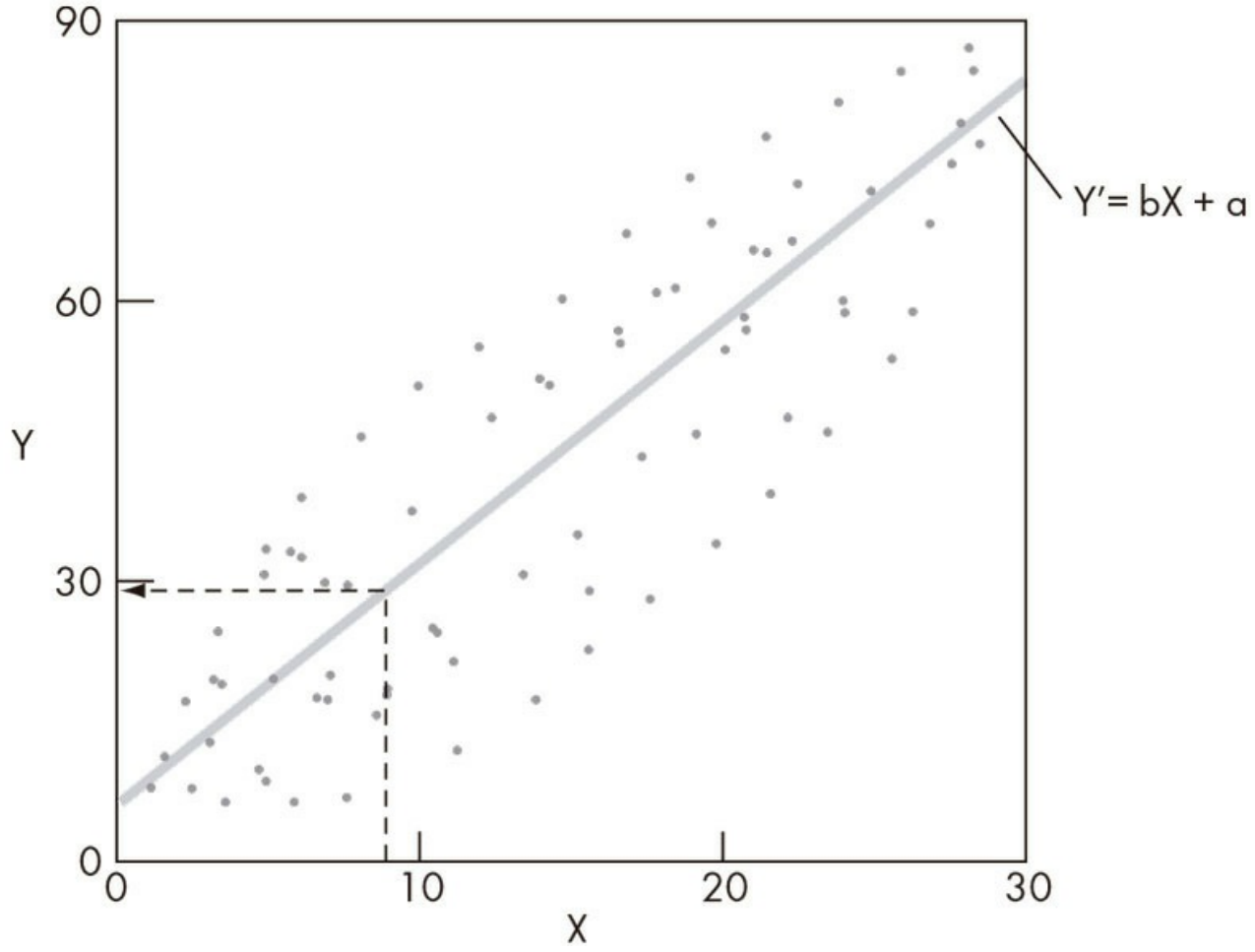


Figura 4-2. Línea de regresión para predecir *Y* a partir de *X*.

Desde luego, no todas las puntuaciones reales de *Y* coinciden de manera exacta con la línea de predicción (a menos que $r = +1.00$ o -1.00). Entonces, hay cierta dispersión de estas puntuaciones de *Y* alrededor de la línea. Mientras más alta es *r*, menor es la dispersión, y mientras más baja es *r*, mayor es la dispersión. Pensemos en la dispersión en un punto de la línea de predicción, en un valor específico de *X*. Suponemos que la distribución de las puntuaciones *Y* es normal; de hecho, asumimos distribuciones normales equivalentes a las puntuaciones *Y'* por cada valor de *X* a lo largo de toda la línea de predicción. La distribución de la figura 4-3 ilustra esta situación. La distribución tiene una desviación estándar, a la cual llamamos **error estándar de estimación** o error estándar de predicción. Usando las características de la distribución normal –por ejemplo, el hecho de que 68% de los casos se encuentran dentro de +/−1 desviación estándar–, podemos hacer afirmaciones acerca de la probabilidad de que las puntuaciones reales difieran de las predichas en cierta magnitud. La fórmula para el error estándar de estimación es:

$$EE_{Y'} = DE_Y \sqrt{1 - r_{xy}^2}$$

Fórmula 4-4

donde DE_Y es la desviación estándar de la prueba que estamos prediciendo y r_{xy} es la correlación entre la prueba que se intenta predecir y la prueba a partir de la cual se hacen las predicciones.

¡Inténtalo!

Para asegurarnos de que entendiste cómo calcular Y', sustituye los siguientes valores en la fórmula y obtén Y'. Estás prediciendo GPA (Y) a partir de las puntuaciones del SAT (X). Para el SAT, la media es 500 y la DE, 100. En el caso del GPA, la media es 2.80 y la DE, .50. La correlación ente GPA y SAT es .65. ¿Qué GPA predice una puntuación de 650 en el SAT?

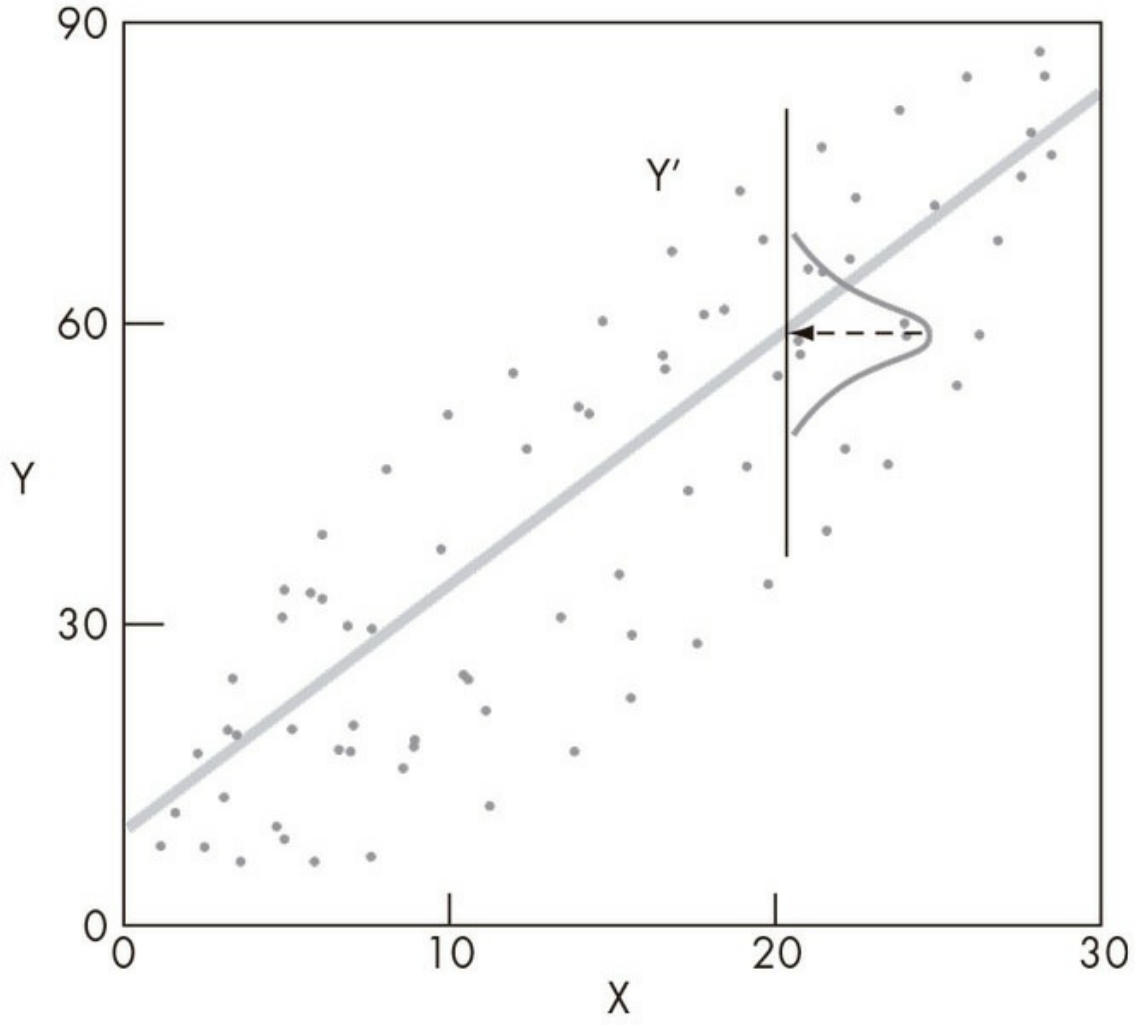


Figura 4-3. Distribución de puntuaciones reales de *Y* alrededor de *Y'*.

Factores que afectan los coeficientes de correlación

Necesitamos tener presentes varios factores que afectan la magnitud de las correlaciones y, por tanto, su interpretación. Primero, el coeficiente de correlación de Pearson, el cual es por mucho el más usado, explica sólo el grado de la relación **lineal** entre dos variables. Si hay cierto grado de no linealidad, la correlación de Pearson subestimará el verdadero grado de la relación. La figura 4-4 muestra una distribución bivariada que tiene cierto grado de curvilinealidad. Una correlación de Pearson explicará la parte lineal de la relación, como lo muestra la línea recta, pero no explicará la tendencia no lineal que se muestra con la línea curveada.

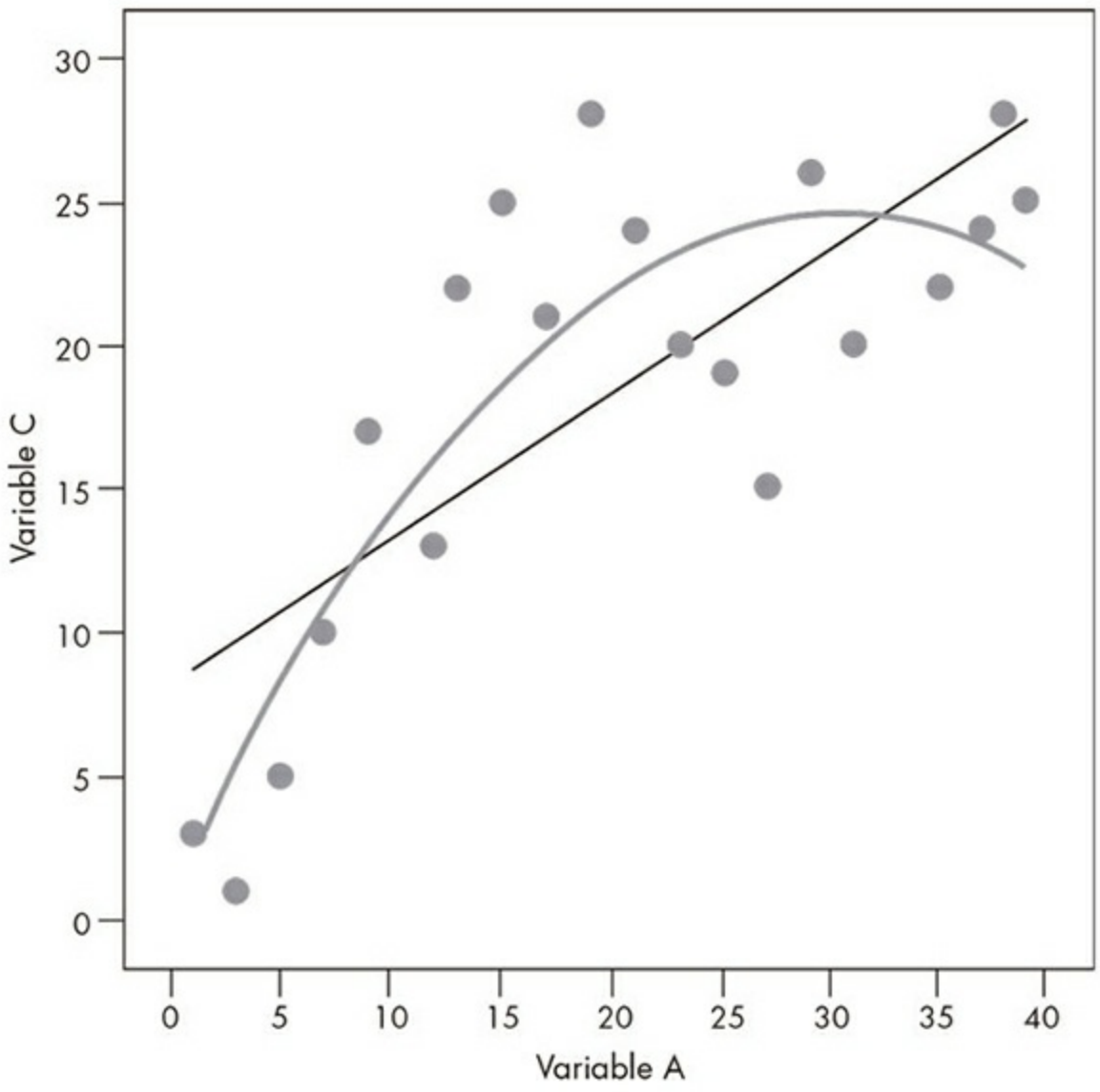


Figura 4-4. Distribución bivariada que muestra una relación curvilínea.

Segundo, como señalamos, suponemos que las puntuaciones *Y* tienen una distribución normal alrededor de la puntuación predicha *Y'* y que el grado de dispersión es igual para cualquier punto a lo largo de la línea de predicción. Esto se conoce como el supuesto de **homocedasticidad** (en griego, de igual dispersión). Sin embargo, es posible que la distribución bivariada muestre **heterocedasticidad** (dispersión diferente), como se muestra en la figura 4-5. Podemos notar que los puntos de los datos se agrupan de manera más estrecha en la parte inferior de la distribución; en cambio, se dispersan más en la parte alta de la distribución. En este caso, el error estándar no es igual en todo el rango de las variables, aunque lo calculamos como si lo fuera.

Tercero, la correlación es estrictamente una cuestión de la posición relativa dentro de cada grupo, de modo que no requiere ni implica puntuaciones absolutas iguales.

Consideremos los datos del cuadro 4-2; si obtenemos las correlaciones entre las puntuaciones de estas pruebas de inteligencia de 10 casos, encontraremos que la correlación entre las pruebas A y B es casi perfecta y sus medias son iguales. En el caso de las pruebas B y C, la correlación es la misma, pero sus medias difieren por 10 puntos. Podríamos inclinarnos a decir que la prueba C no se correlaciona muy bien con la prueba B; sin embargo, $r_{AB} = r_{BC} = .94$. Las posiciones relativas de los casos son iguales entre A y B, y entre B y C aun cuando las puntuaciones absolutas sean superiores en C.

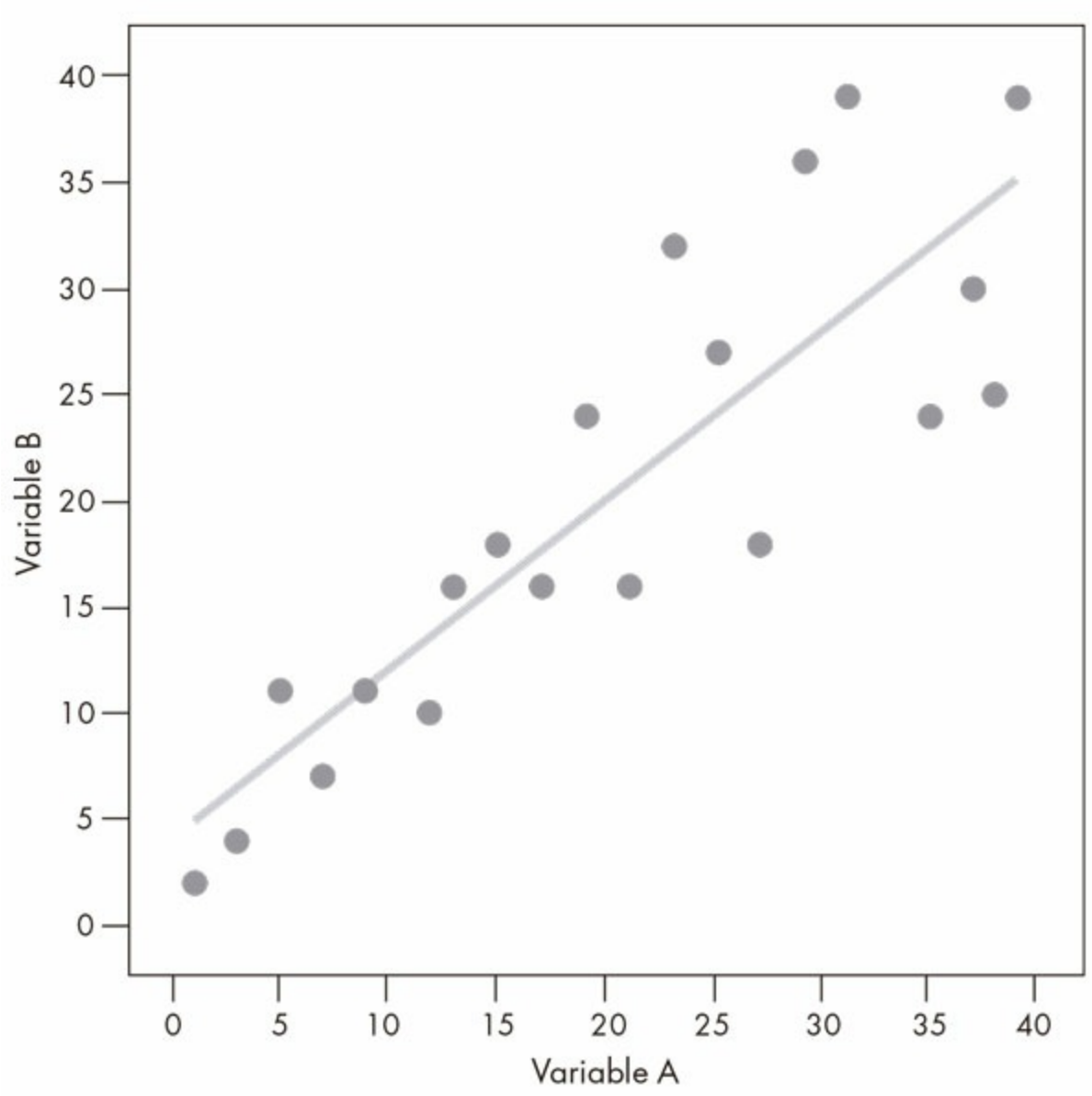


Figura 4-5. Distribución bivariada que presenta heterocedasticidad.

Cuadro 4-2. La correlación es cuestión de posición relativa, no de puntuación absoluta										
Caso	1	2	3	4	5	6	7	8	9	10

Prueba A	80	85	92	90	86	96	100	105	107	110
Prueba B	80	86	85	92	90	96	105	100	107	110
Prueba C	90	95	102	100	96	106	110	115	117	120
$r_{AB} = .94$		$r_{BC} = .94$								
$M_A = 95.1$		$M_B = 95.1$		$M_C = 105.1$						

Cuarto, consideremos el efecto de la variabilidad grupal en el coeficiente de correlación. La desviación estándar o varianza define la variabilidad de un grupo; en este contexto, la variabilidad a menudo se denomina **heterogeneidad** (diferencia), mientras que su opuesto se denomina **homogeneidad** (igualdad). Un grupo muy heterogéneo produce una correlación inflada y uno muy homogéneo, correlaciones reducidas. Consideremos los datos que se muestran en la figura 4-6; si calculamos la *r* del grupo más heterogéneo incluido en el marco A, obtendremos una *r* muy alta. Si hacemos lo mismo con el grupo más homogéneo incluido en el marco C, obtendremos una *r* mucho menor. De los casos en el marco B, obtenemos un valor intermedio de *r*. El ejemplo de la figura 4-6 es un poco artificial, porque implica restringir el rango al mismo.

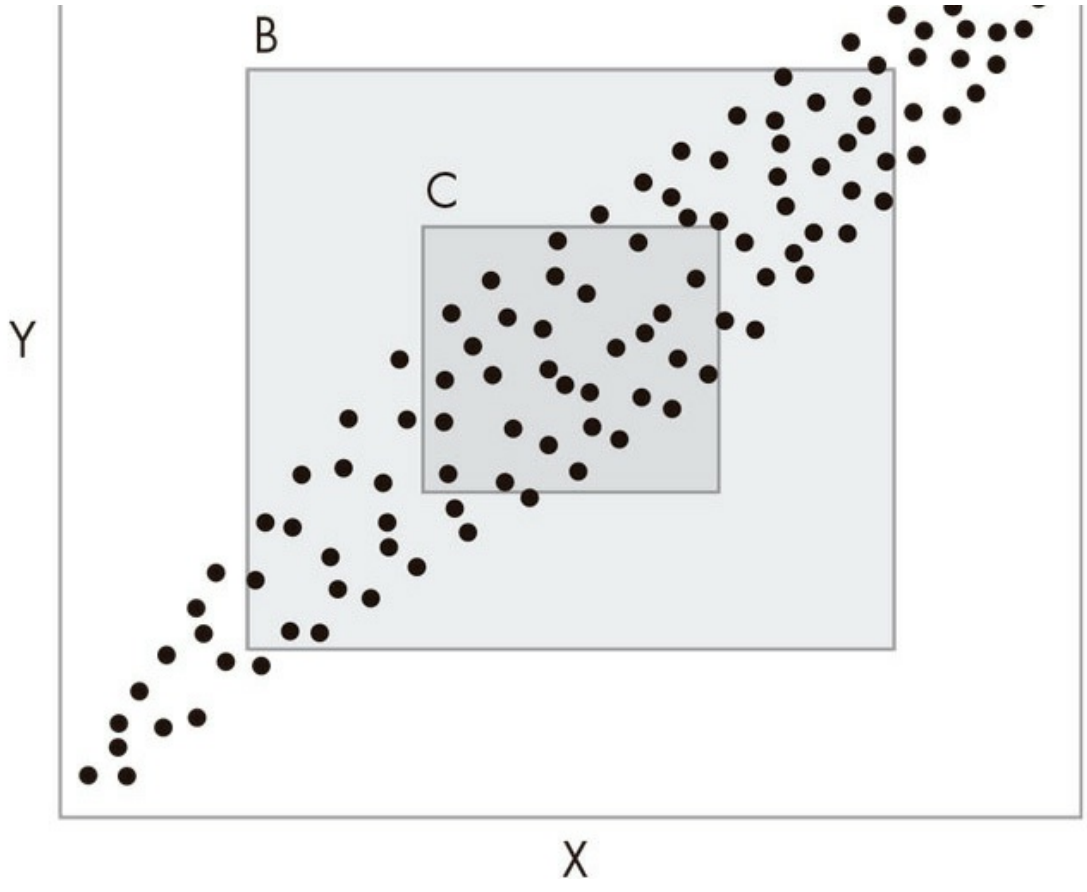


Figura 4-6. Ejemplo del efecto de la restricción del rango en el coeficiente de correlación.

Resumen de puntos clave 4-2

Cuatro factores que afectan el coeficiente de correlación

1. Linealidad
2. Heterocedasticidad
3. Posición relativa (no absoluta)
4. Heterogeneidad grupal

«81-83a

¿Cómo se podrían presentar estas situaciones en la práctica? Supongamos que calculamos la correlación entre las puntuaciones del SAT y del GPA sólo en estudiantes que se han graduado con los máximos honores, es decir, con GPA superiores a 3.90. Este grupo es muy homogéneo con respecto al GPA, por lo que es probable que obtengamos una correlación muy baja, quizá casi de cero, entre el SAT y el GPA en este grupo. Sin embargo, sería temerario concluir que, por regla general, el GPA no está relacionado con las puntuaciones del SAT. Si calculamos la correlación entre el SAT y el GPA de estudiantes de todo el espectro de puntuaciones del GPA, obtendremos un resultado muy diferente. O supongamos que calculamos la correlación entre la talla de zapatos y la puntuación de una prueba de lectura para niños de los grados 1 al 8, un grupo muy heterogéneo. Podríamos obtener una correlación mucho más alta que si hubiéramos limitado la correlación de la talla de los zapatos y la prueba de lectura sólo a niños de octavo grado.

Las diferencias en la variabilidad grupal pueden tener una influencia considerable en la magnitud de la correlación. Necesitamos estar siempre pendientes de esta influencia. Hay fórmulas que permiten la corrección de *r* si la variabilidad del grupo ha aumentado o se ha reducido; se denominan "correcciones por restricción de rango". Aquí proporcionamos las fórmulas más usadas para tales correcciones. El lector que desee conocer una multitud de variaciones de las fórmulas para casos especializados puede consultar Sackett y Yang (2000). Supongamos que conocemos las varianzas (o desviaciones estándar) de los grupos con variabilidad restringida y variabilidad no restringida (o que conocemos una de las varianzas y podemos hacer una estimación razonable de la otra). Luego, conociendo la *r* para el grupo restringido, podemos estimar qué *r* habría en el grupo no restringido. Asimismo, conociendo la *r* del grupo no restringido, podemos estimar qué *r* habría en el grupo restringido. Para estimar qué *r* podría haber en un grupo más heterogéneo a partir de una *r* obtenida de un grupo más homogéneo, usamos la fórmula 4-5. Para estimar qué *r* habría en un grupo más homogéneo a partir de una *r* obtenida de un grupo más heterogéneo, usamos la fórmula 4-6. En Glass y Hopkins (1996) aparecen más ejemplos de los efectos de la restricción de rango sobre el coeficiente de correlación.

$$r_{Me} = \frac{r_{Ma}(DE_{Me} / DE_{Ma})}{\sqrt{1 - r_{Ma}^2 + r_{Ma}^2 (DE_{Me} / DE_{Ma})^2}}$$

Fórmula 4-5

$$r_{Ma} = \frac{r_{Me}(DE_{Ma} / DE_{Me})}{\sqrt{1 - r_{Me}^2 + r_{Me}^2 (DE_{Ma} / DE_{Me})^2}}$$

Fórmula 4-6

r_{Me} = correlación en el grupo *menos* restringido
r_{Ma} = correlación en el grupo *más* restringido
DE_{Me} = desviación estándar del grupo *menos* restringido
DE_{Ma} = desviación estándar del grupo *más* restringido

Para los mortales ordinarios, examinar las fórmulas 4-5 y 4-6 no aporta nada significativo en cuanto al efecto de la homogeneidad grupal en el coeficiente de correlación (*r*). Sin embargo, es más fácil si insertamos valores en las fórmulas y observamos los resultados. Ahora trabajemos con la fórmula 4-5; ésta es la que más se emplea, ya que los proyectos de investigación usan grupos que son *más homogéneos* que la población general. Queremos saber cómo puede cambiar *r* en nuestro proyecto si llevamos a cabo un estudio con poblaciones enteras. Fijemos los valores de *r* en .20, .50 y .90; luego, fijemos la *DE* del grupo más restringido al 50%, 70% y 90% de la *DE* del grupo menos restringido. Por ejemplo, digamos que la *DE* del grupo menos restringido es 10 y que la *DE* del grupo más restringido asume los valores 5, 7 y 9, sucesivamente. Por último, aplicamos la fórmula 4-5 para determinar qué *r* habrá en el grupo menos restringido. El cuadro 4-3 muestra los resultados.

Cuadro 4-3. Valores de muestra para aplicar la corrección por homogeneidad grupal (fórmula 4-5)

DE en un grupo más restringido como porcentaje de la *DE* en el grupo menos restringido	*r* en un grupo más restringido	*r* en un grupo menos restringido
50%	.20	.38
50%	.50	.76
50%	.90	.97
70%	.20	.28
70%	.50	.64
70%	.90	.95
90%	.20	.22

90%	.50	.54
90%	.90	.92

Con base en los datos del cuadro 4-3, así como de simulaciones adicionales, concluimos lo siguiente. Primero, la corrección para la heterogeneidad del grupo tiene efectos considerables cuando la variabilidad es mucho menor en el grupo restringido que en el grupo no restringido. Por ejemplo, cuando la *DE* en el grupo restringido es sólo la mitad de la *DE* del grupo no restringido, *r* puede aumentar en más de 20 puntos. Sin embargo, cuando la *DE* del grupo restringido alcanza 90% de la *DE* del grupo no restringido, el efecto sobre *r* es mínimo. Segundo, el efecto de la corrección es más pronunciado con niveles moderados de correlación. Ésta es una conclusión importante, porque la mayoría de las correlaciones con que trabajamos en psicología son moderadas. Correlaciones muy bajas (p. ej., debajo de .10) y muy altas (p. ej., arriba de .90) son poco afectadas por la corrección para la homogeneidad grupal. En caso de que esto no sea evidente de inmediato, debemos señalar que la corrección nunca resulta en un cambio de la dirección de la relación.

Principales fuentes que atentan contra la confiabilidad

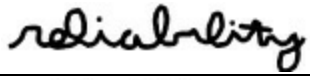
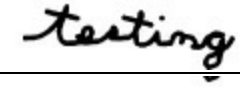
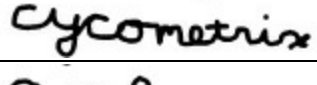
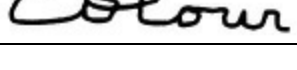
Antes de formular los métodos específicos para expresar la confiabilidad de las pruebas, es importante considerar las fuentes potenciales que atentan contra la confiabilidad. ¿Qué factores o condiciones llevarán a una medición menos confiable? Justo estos factores son los que los índices de confiabilidad deben abordar. Cualquier cosa que resulte en una variación no sistemática de las puntuaciones de la prueba es una fuente de falta de confiabilidad. Ninguna lista puede ser exhaustiva, por lo que aquí identificamos cuatro categorías principales de estas fuentes.

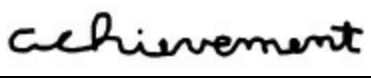
Calificación de la prueba

La variación en la calificación de la prueba, como una fuente que afecta a la confiabilidad, es una de las más fáciles de entender, pero también es de gran importancia histórica. La preocupación por las diferencias en las puntuaciones de un juez a otro –incluso en pruebas sencillas como las de ortografía o cálculo aritmético– fue una fortaleza importante en el desarrollo de los reactivos de opción múltiple para las pruebas de aprovechamiento y de capacidad.

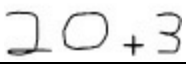
Consideremos los casos sencillos que se presentan en los cuadros 4-4 y 4-5. En el caso de las respuestas de una prueba de ortografía del inglés resumidas en el cuadro 4-5, ambos jueces, 1 y 2, concuerdan en que las dos primeras palabras están escritas correctamente y que la tercera no lo está. Sin embargo, el juez 1 concede crédito a "colour" como una variante legítima de "color", mientras que el juez 2 no lo hace. En cuanto a "achievement", el juez 1 le concede al estudiante el "beneficio de la duda" por las ambiguas "ie" a mitad de la palabra, pero el juez 2 no es tan amable. Así, la puntuación de estos cinco reactivos varía en dos puntos (¡40%!) dependiendo de quién calificó las respuestas.

Cuadro 4-4. Respuestas a la prueba dictada de ortografía de inglés

Palabra dictada	Respuesta del estudiante	Juez 1	Juez 2
reliability		C	C
testing		C	C
psychometrics		I	I
color		C	I
achievement		C	I

			
Puntuación total		4	2
C = correcta, I = incorrecta.			

Cuadro 4-5. Respuestas de la prueba de cálculo aritmético

Reactivo	Respuesta del estudiante	Juez 1	Juez 2
6 + 2		C	C
10 − 5		C	C
3 × 3		I	I
4 + 3		C	I
35 − 12		C	I
Puntuación total		4	2
C = correcta, I = incorrecta.			

El cuadro 4-5 muestra ejemplos de respuestas de una prueba de cálculo aritmético sencillo. Los jueces 1 y 2 concuerdan en que las respuestas a los primeros dos reactivos son correctas y que la del tercero es incorrecta.

Sin embargo, en el tercer reactivo, el juez 1 concede crédito al 7 invertido, pues nota que el estudiante obviamente sabe el resultado de la operación, pero tuvo dificultades para escribir el número de la manera correcta. El juez 2 insiste en que el resultado esté expresado correctamente. En el quinto reactivo, el juez 1 nota con generosidad que la respuesta del estudiante es técnicamente correcta aunque no esté expresada de una manera estándar. El juez 2, adusto, encuentra esta expresión por completo inaceptable.

Estos ejemplos muestran cómo las variaciones en los criterios de calificación pueden afectar incluso los reactivos más sencillos. Consideremos cuánta variación puede encontrarse al calificar respuestas a reactivos como las preguntas abiertas en una prueba de inteligencia de aplicación individual, una escala para valorar creatividad o una prueba proyectiva de personalidad. Por ejemplo, muchas pruebas de inteligencia incluyen reactivos de vocabulario; el examinador dice una palabra y el examinado debe dar una definición aceptable. El cuadro 4-6 muestra dos ejemplos de palabras y varias respuestas para cada una; cada respuesta se califica con 0 (claramente incorrecta), 1 (parcialmente correcta) o 2 (claramente correcta).

Cuadro 4-6. Respuestas muestra a los reactivos de Vocabulario

Palabra	Respuestas de examinados	Puntuación 0, 1, 2

Confiable	> como, tú sabes, ser consistente, cumplidor	( )
	> duro, difícil	( )
	> ser lo mismo, idéntico	( )
	> usual	( )
Escuela	> un edificio	( )
	> un lugar a donde van los estudiantes a aprender	( )
	> un grupo de tipos	( )
	> un montón de libros	( )
	> donde viven los maestros	( )

¡Inténtalo!

Califica las respuestas del cuadro 4-6. Compara tus calificaciones con las de otros estudiantes.

En resumen, la falta de acuerdo entre los jueces puede resultar en una variación no sistemática en las puntuaciones de las pruebas. Las máquinas que califican reactivos de "opción" por lo general eliminan tal variación, pero ni siquiera ellas están por completo exentas de errores. Mientras mayor criterio se requiere para calificar, confiabilidad. Cuando se requiere un criterio para calificar una prueba,, la meta es tener instrucciones de calificación que sean suficientemente claras y explícitas para que la variación debida a los jueces se reduzca al mínimo.

Contenido de la prueba

Las variaciones en el muestreo de los reactivos de una prueba pueden resultar en un error no sistemático en las puntuaciones. Consideremos una prueba de matemáticas usada para ubicar estudiantes en los cursos de matemáticas. de una universidad, la cual tiene 10 versiones ligeramente distintas para usarlas con los alumnos de nuevo ingreso a lo largo de las sesiones de orientación de verano. Una versión tiene dos reactivos sobre el teorema de Pitágoras, mientras que otra versión sólo tiene un reactivo sobre este tema. Un estudiante que tenga un particular dominio del teorema puede obtener una puntuación ligeramente más alta en la primera versión que en la segunda. O consideremos a dos estudiantes preparándose para un examen de historia; el examen abarcará seis capítulos. La profesora incluirá en la prueba cuatro preguntas abiertas de un número potencialmente infinito que tiene en mente. Un estudiante se concentra en los primeros cuatro capítulos y da un repaso superficial a los otros dos. Otro estudiante lee rápido los primeros dos capítulos y se concentra en los últimos cuatro. Para el examen, los dos estudiantes saben la misma cantidad de material; sin embargo, tres de las cuatro preguntas provienen de los últimos cuatro capítulos. ¿Cómo afecta la variación del contenido las puntuaciones de ambos estudiantes? ¿Qué pasaría si tres de las cuatro

preguntas fueran tomadas de los primeros cuatro capítulos?

Estas ligeras variaciones en el muestreo de los reactivos de una prueba producen errores no sistemáticos. Las puntuaciones de los individuos aumentan o disminuyen, quizá sólo por pocos puntos, quizá por más, pero no a causa de diferencias reales en el rasgo que se mide, sino debido a cambios más o menos aleatorios en el conjunto de reactivos que constituyen la prueba.

Condiciones de aplicación de la prueba

Una prueba debe tener procedimientos estandarizados para su aplicación, los cuales incluyen instrucciones, límites de tiempo y condiciones físicas del lugar. Sin embargo, es imposible controlar todos los detalles imaginables de la aplicación aun sabiendo que tendrán alguna influencia en las puntuaciones finales. Por ejemplo, el ruido de una avenida o iluminación insuficiente durante la aplicación pueden afectar de manera negativa la puntuación de la prueba. Si una prueba tiene un límite de tiempo de 30 minutos, un aplicador puede ser un poco más generoso y conceder quizá 31 minutos, mientras que otro puede ser bastante estricto y dar 29.5 minutos. Todas estas pequeñas variaciones en la aplicación de la prueba pueden ser fuentes de varianza inestable en las puntuaciones.

Condiciones personales

Las condiciones temporales del examinado pueden tener influencias no sistemáticas en sus puntuaciones. Si se le aplica la prueba el martes, Luis puede obtener una puntuación algo inferior porque está un poco resfriado. Si la prueba fuera el miércoles, cuando se sienta mucho mejor, podría obtener algunos puntos extra. Jen está de pésimo humor el viernes, cuando le aplicaron un inventario de personalidad; si se lo hubieran aplicado el sábado, cuando ya estaba más relajada, su puntuación habría sido diferente. En ambos casos, no hay diferencia de un día a otro en el rasgo subyacente que se mide, pero la situación personal influye en las puntuaciones.

Las variaciones en los factores que hemos considerado no dan por resultado, de manera automática, falta de confiabilidad; por ejemplo, variaciones en la iluminación del cuarto o un resfriado sin importancia pueden no afectar el desempeño en la prueba. El grado en que estos factores afectan la puntuación es una cuestión empírica, la cual abordaremos en la siguiente sección al considerar los métodos con que se determina y expresa la confiabilidad de la prueba. Ahora trataremos formalmente estos métodos.

Marco conceptual: teoría de la puntuación verdadera

La confiabilidad de las pruebas puede formularse dentro de tres contextos teóricos en cierto modo diferentes: teoría clásica de las pruebas (TCP), teoría de la respuesta al reactivo (TRR) y teoría de la generalizabilidad (TG). La gran mayoría de la información sobre confiabilidad que encontramos actualmente en los manuales de las pruebas, revistas científicas e informes de evaluación se apoya en la TCP. Por ello, en este capítulo nos concentramos en esta teoría; sin embargo, la TRR y la TG están ganando popularidad, así que las presentamos al final de este capítulo.

La teoría clásica de las pruebas comienza con un marco conceptual interesante y útil. Las palabras clave de este marco son *puntuación observada* (*O*), *puntuación verdadera* (*V*) y *puntuación de error* (*E*). La **puntuación observada** es la puntuación real de una persona en una prueba; podemos pensarla como la puntuación natural –por ejemplo, 30 reactivos correctos de 45 en una prueba de solución de problemas aritméticos– aunque el concepto se aplica igual de bien a las puntuaciones normativas tales como las estándar. La puntuación observada puede ser afectada, en sentido positivo o negativo, por varias fuentes que afectan a la confiabilidad; por ejemplo, esta puntuación puede ser un poco alta debido a la buena suerte al responder preguntas de las que no se conoce la respuesta correcta, o puede ser un poco baja debido a que el examinado estuvo demasiado cansado durante la aplicación.

La **puntuación verdadera** es la que una persona obtendría si todas las fuentes que afectan a la confiabilidad pudieran ser eliminadas o canceladas. Podríamos pensarla como la puntuación promedio obtenida en muchas aplicaciones (en teoría, un número infinito de ellas) de la prueba en distintos momentos y en condiciones ligeramente diferentes. Cada variación en la aplicación puede introducir cierta falta de confiabilidad, pero cuando todas las puntuaciones reales u observadas se promedian, la media podría ser igual a la puntuación verdadera. Ésta es la que en realidad queremos conocer, aunque en la práctica nunca podemos estar por completo seguros, pues sólo tenemos una puntuación observada.

> Decir que una puntuación incluye error implica que hay un valor hipotético libre de errores que caracteriza la variable que se evalúa. En la teoría clásica de las pruebas, este valor se denomina *puntuación verdadera* de la persona en la prueba. Se conceptualiza como el promedio hipotético de puntuaciones obtenidas de una serie infinita de réplicas del procedimiento de aplicación.
>
> *Standards*... (AERA, APA, & NCME, 2013)

La **puntuación de error** es sólo la diferencia entre la puntuación verdadera y la puntuación observada. *E* puede ser positivo o negativo. Es la sumatoria de todas las influencias no sistemáticas en la puntuación real de una persona que abordarnos en la sección de factores que atentan contra la confiabilidad. La fórmula 4-7 expresa las

relaciones entre las puntuaciones observada, verdadera y de error.

$$V = O \pm E$$

Fórmula 4-7

La fórmula también podría escribirse como:

$$O = V \pm E$$

Fórmula 4-8

o

$$\pm E = V - O.$$

Desde luego, las tres fórmulas son equivalentes en términos algebraicos, pero cada una ofrece un modo ligeramente distinto de pensar la relación. Podemos notar que la puntuación de error puede ser positiva o negativa.[2]

La teoría de la puntuación verdadera también puede expresarse en términos de las varianzas de las puntuaciones de la prueba. Recordemos que la varianza es sólo la desviación estándar elevada al cuadrado. En esta formulación,

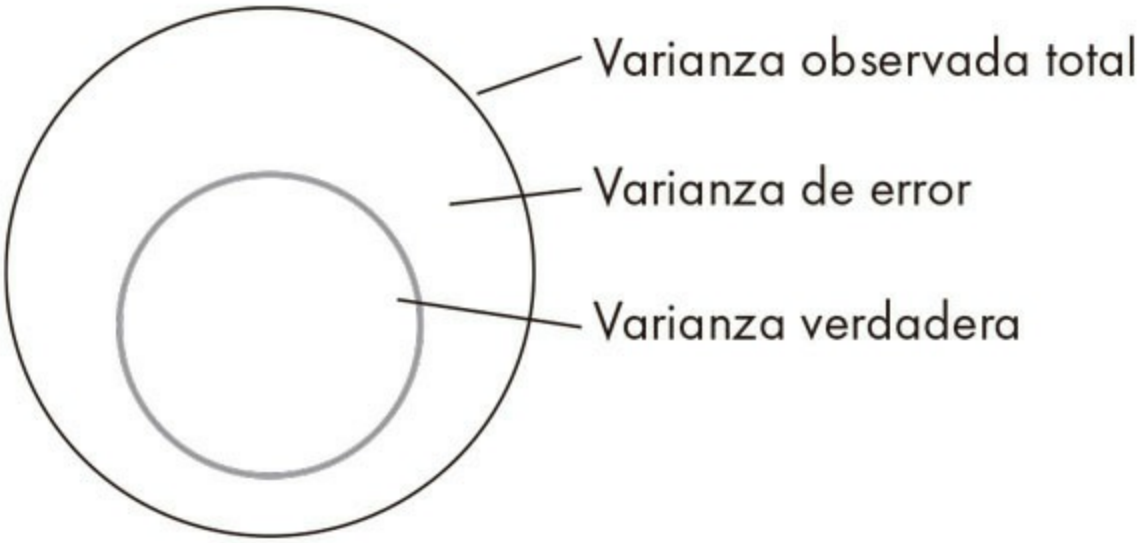
$$\sigma^2_O = \sigma^2_V + \sigma^2_E$$

Fórmula 4-9

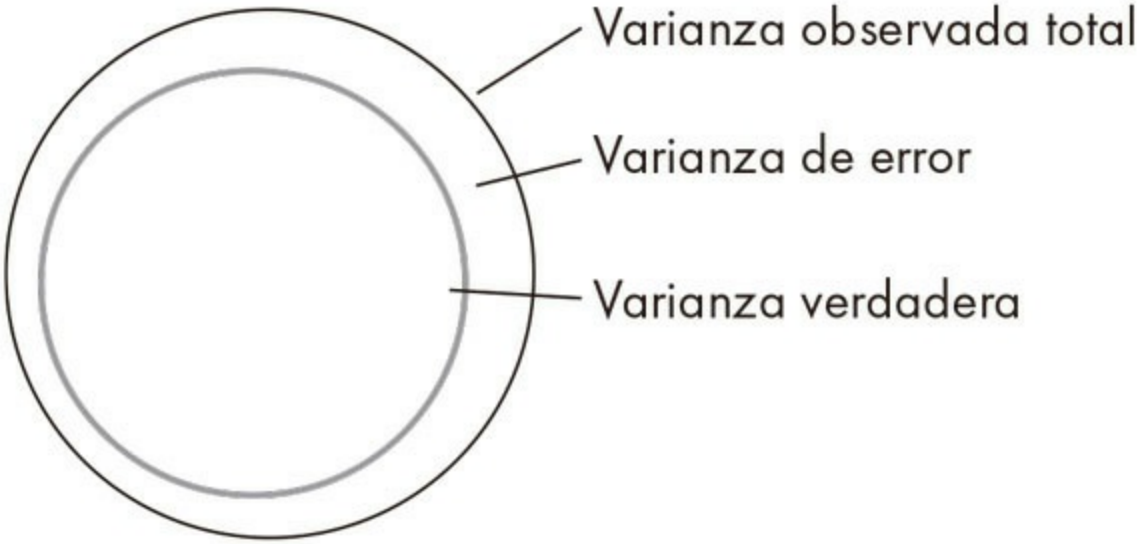
o

$$\sigma^2_V = \sigma^2_O - \sigma^2_E$$

Es decir, la varianza de las puntuaciones observadas es la suma de la varianza de la puntuación verdadera y la varianza de la puntuación de error, o bien, la varianza de la puntuación verdadera es igual a la varianza observada menos la varianza de error. Estas relaciones se representan en la figura 4-7. El panel A muestra una prueba en la que la varianza verdadera representa sólo la mitad de la varianza observada; el resto es varianza de error. El panel B muestra una prueba en la que la varianza de error es una fracción relativamente pequeña del total de la varianza observada; por lo que la mayor parte es varianza verdadera. En otras palabras, la prueba del panel B tiene una confiabilidad mucho mejor que la del panel A.



Panel A



Panel B

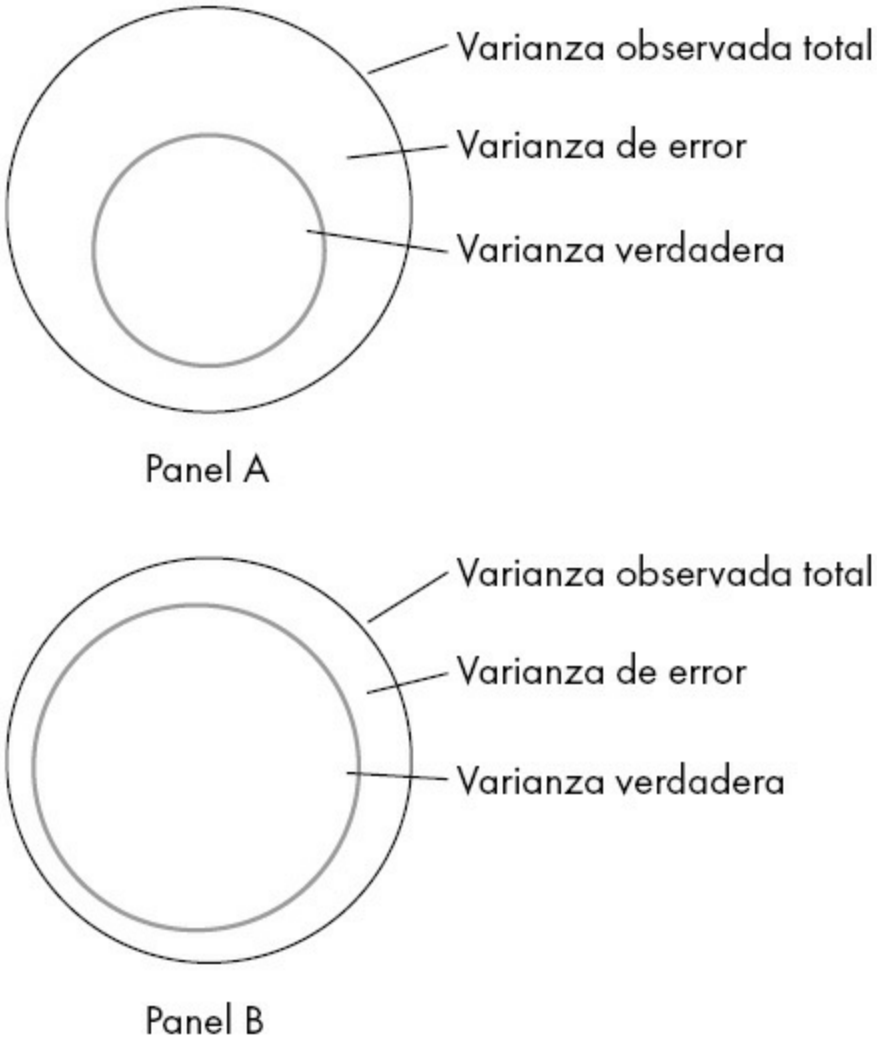


Figura 4-7. Relaciones ente las varianzas observada, verdadera y de error.

Con base en la notación adoptada aquí, podemos definir la confiabilidad (*r*) como:

$$r = \frac{\sigma_V^2}{\sigma_O^2}$$

Fórmula 4-10

es decir, como la proporción de varianza observada de la puntuación que es varianza verdadera. Otra forma en que se puede presentar esta última fórmula es:

$$r = \frac{\sigma_O^2 - \sigma_E^2}{\sigma_O^2}$$

Fórmula 4-11

Esta fórmula será importante en algunos tratamientos más avanzados de la confiabilidad.

Como sugerimos antes, es conveniente pensar en la puntuación verdadera de una persona como el promedio de muchas puntuaciones observadas. La figura 4-8 muestra ejemplos de distribuciones que resultan de muchas aplicaciones de dos pruebas. En la gráfica A, la prueba es muy confiable, pues las puntuaciones observadas se agrupan de manera estrecha alrededor de la puntuación verdadera *V*. En la gráfica B, la prueba no es muy confiable, pues las puntuaciones observadas se dispersan ampliamente alrededor del promedio o puntuación verdadera *V*. La diferencia entre cualquier puntuación *O* y *V* en esta distribución es error de medición *E*. Solemos suponer que las puntuaciones observadas tienen una distribución normal alrededor de la puntuación verdadera. La figura 4-8 parte de este supuesto, lo que tendrá consecuencias convenientes más adelante en este capítulo (véase Error estándar de medición, 93a»).

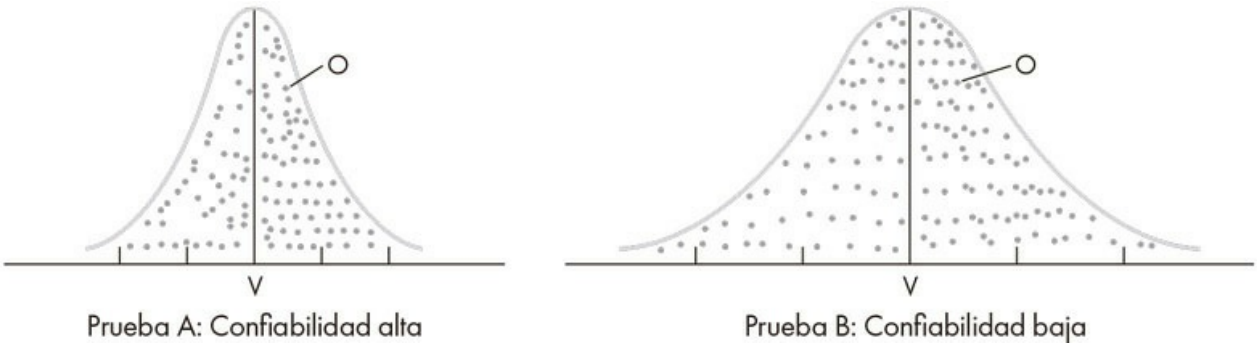


Figura 4-8. Distribuciones hipotéticas de puntuaciones observadas (*O*) alrededor de puntuaciones verdaderas (*V*).

Por lo común, en situaciones prácticas de evaluación, obtenemos *sólo una* puntuación observada, de modo que la distribución de puntuaciones observadas, como la que aparece en la figura 4-8, es meramente hipotética. Ése es el modo en que las puntuaciones observadas se distribuirían, suponemos, *si* obtuviéramos muchas de ellas de un solo individuo. Nuestro estudio de la confiabilidad ayudará a operacionalizar este supuesto.

Como señalamos antes, nunca podemos saber la puntuación verdadera de una persona aunque eso es lo que queremos. Siempre tenemos sólo una puntuación observada. Los distintos métodos para determinar la confiabilidad, que revisaremos a continuación, están diseñados para estimar qué tanta diferencia puede haber entre la puntuación observada y

la puntuación verdadera de una persona, es decir, cuánto error puede haber en la medición.

Métodos para determinar la confiabilidad

Pueden usarse diferentes métodos para determinar la confiabilidad de una prueba; cada uno de ellos trata una o más de las fuentes que afectan a la confiabilidad y que describimos antes. Aquí consideraremos los métodos que se utilizan con mayor frecuencia. A lo largo de toda esta sección, nos situaremos dentro del marco conceptual de la teoría clásica de las pruebas.

Confiabilidad de test-retest

Uno de los tipos más fáciles de entender es la **confiabilidad de test-retest**. Como lo sugiere su nombre, este coeficiente de confiabilidad se obtiene aplicando la misma prueba a los mismos individuos en dos ocasiones diferentes, que, por lo general, pueden estar separadas desde un día hasta un mes. Este coeficiente consiste simplemente en la correlación (casi siempre la de Pearson) entre las puntuaciones de la primera aplicación y las de la segunda. A menudo se le llama coeficiente de estabilidad temporal. El cuadro 4-7 presenta un conjunto de datos que pertenecen a un estudio de confiabilidad de test-retest.

Cuadro 4-7. Datos para determinar la confiabilidad de test-retest

Examinado	Primera aplicación	Segunda aplicación
1	85	81
2	92	79
3	76	75
4	61	69
5	93	93
—	—	—
—	—	—
—	—	—
100	80	82
	r = .85	

¿Qué fuentes de las que describimos antes afectan este tipo de confiabilidad? Es evidente que evaluar la influencia de los cambios en las condiciones personales ayuda; también es evidente que *no* se trata de la influencia de los cambios en el contenido de la prueba, ya que se emplea exactamente la misma. La confiabilidad test-retest puede o puede no relacionarse con variaciones debidas a la aplicación de la prueba, dependiendo de si la prueba es aplicada por la misma persona, en el mismo lugar, etc., en ambas ocasiones. Además, puede o puede no indicar variaciones interjueces, dependiendo de si la prueba es calificada por la misma persona o siguiendo el mismo procedimiento en

ambas ocasiones.

Determinar la confiabilidad mediante este método tiene tres inconvenientes principales. Primero, el método obviamente no toma en cuenta el error no sistemático debido a las variaciones en el contenido de la prueba. Segundo, en cualquier prueba, excepto las más sencillas y breves, obtener la confiabilidad de test-retest es un fastidio, ¿Quién quiere hacer la misma prueba de cuatro horas dos veces en un periodo de dos semanas? Tercero, existe cierta preocupación acerca del efecto de la primera aplicación en la segunda; quizá el examinado recordará las respuestas de la primera aplicación y, de manera deliberada, dará las mismas respuestas en la segunda buscando ser consistente aun cuando piense de modo diferente. Esto tiende a inflar el coeficiente de confiabilidad. Si un reactivo demanda una solución de problemas novedosa y el examinado falla en la primera aplicación, es posible que más tarde se le ocurra otra solución que le permita responder correctamente ese reactivo en la segunda aplicación. En los reactivos que demandan información, el examinado puede buscar la respuesta correcta entre la primera y la segunda aplicación. El grado en que estos factores pueden influir en las puntuaciones de la segunda aplicación es, en gran medida, cuestión de criterio.

El tiempo entre la primera y la segunda aplicación es motivo de preocupación para la confiabilidad test-retest. Por un lado, los intervalos deben ser suficientemente largos para que la primera aplicación tenga una influencia mínima sobre la segunda. Por otro lado, el intervalo no debe ser tan largo como para que el rasgo que se mide pueda sufrir cambios reales. Por tomar un ejemplo extremo, si el intervalo entre dos aplicaciones fuera de cinco años, podríamos suponer que la correlación entre la primera y la segunda estaría en función de cambios reales en el rasgo más que de la confiabilidad de la prueba. En la práctica, los estudios de confiabilidad de test-retest, por lo general, emplean intervalos de unos cuantos días o semanas. Sin embargo, no hay una regla definida en relación con este asunto.

Resumen de puntos clave 4-4

Métodos para determinar la confiabilidad

- Test-retest
- Interjueces
- Formas alternas
- Consistencia interna

Confiabilidad interjueces

La **confiabilidad interjueces** se puede entender con facilidad. Evalúa la variación no sistemática debida simplemente a quién califica la prueba. “Quién” se refiere, por lo común, a dos personas diferentes, aunque también podría referirse a dos máquinas o a

una persona y una máquina, o cualquier otra combinación. Este tipo de confiabilidad también se podría denominar interobservadores o inter-evaluadores de acuerdo con lo que en realidad se hace en la situación de prueba. Por ejemplo, dos personas pueden valorar la creatividad o la gravedad del desajuste del examinado. También se usan formas alternas para denominar esta confiabilidad: **confiabilidad de jueces, confiabilidad de observadores o confiabilidad de evaluadores.**

La confiabilidad interjueces se obtiene con facilidad. Se aplica una prueba a un grupo y se califica dos veces. La confiabilidad interjueces es simplemente la correlación, casi siempre la de Pearson, entre las calificaciones del primer juez con las del segundo. Los datos del cuadro 4-7 pueden usarse para este tipo de confiabilidad simplemente cambiando el encabezado de las columnas Primera aplicación y Segunda aplicación por Juez A y Juez B.

Es importante que los dos (o más) jueces trabajen de manera independiente, es decir, ninguno de ellos debe influir en el otro o los otros jueces. Por ejemplo, si el juez B sabe qué puntuación asignó el juez A a un reactivo o a la prueba entera, podría inclinarse a asignar la misma calificación o una parecida, lo cual inflaría el coeficiente de confiabilidad resultante. Desde luego, también podríamos imaginar que el juez B detesta al juez A, por lo que estará en desacuerdo de manera deliberada al asignar calificaciones, lo cual disminuiría el coeficiente de confiabilidad. Cualquiera que sea el caso, la influencia de un juez sobre otro contamina el estudio de la confiabilidad, por lo que los jueces deben trabajar de manera independiente.

En algunos estudios se requiere de más de dos jueces; por ejemplo, después de la entrevista inicial con 50 clientes, cuatro clínicos de manera independiente valoran el grado de desajuste en una escala de 20 puntos que va de "gravemente desajustado" a "sin desajuste perceptible". El cuadro 4-8 presenta las valoraciones de algunos clientes. ¿Cómo se expresa el acuerdo interjueces en esta situación? Sería posible calcular las correlaciones entre todas las posibles combinaciones de jueces (A y B, A y C, A y D, B y C, B y D, C y D) y, luego, promediar las correlaciones. De hecho, esto se hace a veces; sin embargo, el análisis más apropiado para esta situación es el **coeficiente de correlación intraclase**, que se abrevia como r_I o ρ_I en los libros de estadística, pero en otras fuentes se escribe simplemente CCI. El CCI se calcula a partir de las medias cuadráticas (MC) desarrolladas en el análisis de varianza (ANOVA). Existe una sorprendente cantidad de maneras para definir y calcular el CCI, que se pueden consultar en Shrout y Fleiss (1979) o Winer (1991). Para nuestros propósitos, el punto importante es que el CCI se interpreta como el conocido coeficiente de correlación de Pearson (*r*). Tratándose de confiabilidad, el objetivo de la aplicación usual del CCI es determinar la confiabilidad interjueces.

Cuadro 4-8. Datos para estudiar la confiabilidad interjueces con más de dos jueces

Cliente	**Clínico**			
	A	**B**	**C**	**D**

	Valoraciones de desajustes			
1	15	12	13	14
2	8	7	7	6
3	12	18	8	10
4	14	10	14	9
—	—	—	—	—
—	—	—	—	—
50	6	4	5	3

La confiabilidad interjueces obviamente proporciona información sobre los errores no sistemáticos que surgen de los distintos jueces, pero de ninguna otra fuente de error. La información de este tipo de confiabilidad es de particular importancia cuando el juicio de los evaluadores interviene en el proceso de calificación.

Confiabilidad de formas alternas «89-93a

También conocida como confiabilidad de formas paralelas o equivalentes, la **confiabilidad de formas alternas** requiere que existan dos formas de la prueba. Éstas deben ser iguales o muy similares en términos del número de reactivos, límites de tiempo, especificaciones del contenido y de otros factores de este tipo.[3] El estudio de la confiabilidad de formas alternas consiste en aplicar ambas formas de la prueba a los mismos examinados. Este tipo de confiabilidad es la correlación, casi siempre la de Pearson, entre las puntuaciones obtenidas en las dos formas de la prueba. Regresemos al cuadro 4-7; si cambiamos los encabezados de las columnas Primera aplicación y Segunda aplicación por Forma 1 y Forma 2, tendremos el diseño básico para estudiar la confiabilidad de formas alternas.

Las formas alternas de la prueba pueden aplicarse de manera inmediata una después de la otra si son relativamente breves y poco demandantes; de lo contrario, puede emplearse un intervalo similar al que se usa para obtener la confiabilidad de test-retest. En el caso más sencillo de confiabilidad de formas alternas, cuando las dos formas se aplican una inmediatamente después de la otra, el método sólo mide la falta de confiabilidad debida al muestreo del contenido. Cuando se trata de pruebas más extensas, las formas alternas suelen aplicarse a intervalos de algunos días o algunas semanas, en cuyo caso, el método mide falta de confiabilidad debida al muestreo del contenido y, como en la confiabilidad de test-retest, cambia con las condiciones personales y las variaciones en la aplicación.

La confiabilidad de formas alternas no se usa con mucha frecuencia por la sencilla razón de que la mayoría de las pruebas no tiene formas alternas. Si crear una buena prueba es bastante difícil, crear dos o más formas alternas, más o menos equivalentes, lo es aún más. Las formas alternas suelen estar disponibles sólo en el caso de algunas de las pruebas más usadas.

> Se reconocen tres amplias categorías de coeficientes de confiabilidad: a) coeficientes derivados de la aplicación de formas alternas en sesiones independientes (coeficientes de formas alternas), b) coeficientes obtenidos aplicando la misma forma en ocasiones separadas (coeficientes de test-retest) y c) coeficientes basados en las relaciones /interacciones entre puntuaciones derivadas de reactivos individuales o subconjuntos de ellos dentro de una prueba; todos los datos proceden de una sola aplicación (coeficientes de consistencia interna). Además, cuando la calificación de la prueba implica un elevado uso del de calificación a través de jueces, suelen obtenerse los índices de la consistencia entre jueces.
>
> *Standards*... AERA, APA, & NCME, 2013)

«90c

Confiabilidad de consistencia interna «90b

La **consistencia interna** es uno de los métodos de confiabilidad que se usa con mayor frecuencia. Existen numerosos métodos para determinar la confiabilidad de consistencia interna de una prueba. Describiremos tres de los métodos más usados: división por mitades, Kuder-Richardson y coeficiente alpha. Todos los métodos de consistencia interna, incluyendo los que no revisaremos aquí, intentan medir las características comunes de la consistencia interna de la prueba.

Los métodos de consistencia interna, como otros que hemos considerado en este libro, producen un coeficiente de confiabilidad. Sin embargo, lo que sucede con exactitud con estos métodos es menos evidente que con otros métodos; una vez descritos, los métodos de test-retest, de interjueces y de formas alternas parecen claros a nivel intuitivo. Pero eso no ocurre con los métodos de consistencia interna, por lo que tendremos que comenzar describiendo su lógica; el primero será el método de división por mitades.

Confiabilidad de división por mitades «90a

Recordemos el método de formas paralelas de la sección anterior. Ahora pensemos en el caso específico en que las dos formas se aplican en sucesión inmediata. Pensemos, entonces, en la aplicación de una sola prueba, pero la cual calificaremos por mitades, como si cada una fuera una forma alterna de la prueba. Después correlacionamos las puntuaciones de las dos mitades de la prueba. Esto es como una medida de confiabilidad de "miniformas alternas", que es lo que en esencia sucede con la **confiabilidad de división por mitades**.

Hay dos importantes desarrollos en este último escenario. Primero, la prueba *no* suele dividirse en dos tomando la primera parte y la segunda, porque a menudo en la segunda parte de la prueba se encuentran los reactivos más difíciles. Los examinados pueden estar más fatigados al final de la prueba, y si hay algún efecto del tiempo, es más probable que su influencia sea mayor en la segunda parte que en la primera. Entonces, ¿cómo se divide por la mitad una prueba? Un método que se usa con frecuencia consiste en dividir la prueba en reactivos pares y reactivos nones, en cuyo caso el resultado suele denominarse **confiabilidad de pares y nones**. Otros tipos de división pueden ser útiles con cierta clase de reactivos, pero el método de pares y nones es, por mucho, el más

empleado.

Segundo, la correlación entre las dos mitades de la prueba no indica la confiabilidad de la prueba entera, sino de la mitad de la prueba en que estamos interesados. Por ello, debe aplicarse una corrección a la correlación entre las mitades para obtener la confiabilidad de la prueba entera. La **corrección de Spearman-Brown** es la adecuada, y su fórmula es:

$$r_c = \frac{2r_m}{1 + r_m}$$

Fórmula 4-12

r_c = confiabilidad corregida de toda la prueba
r_m = correlación entre las dos mitades de la prueba

La fórmula de Spearman-Brown tiene una forma más general que permite determinar el efecto estimado sobre la confiabilidad de consistencia interna de cualquier cambio en la extensión de la prueba. La forma más general es:

$$r_c = \frac{nr_o}{1 + (n-1)r_o}$$

Fórmula 4-13

n = factor por el cual se cambia la extensión de la prueba
r_c = confiabilidad corregida
r_o = confiabilidad original

En esta fórmula, n puede ser una fracción; por ejemplo, es posible estimar la confiabilidad corregida de una cuarta parte (n = .25) de la prueba original. También se puede estimar el efecto de triplicar (n = 3) la extensión de la prueba. O se puede fijar r_c en algún valor deseado, y luego encontrar el valor de n para determinar qué cambio en la extensión de la prueba se requiere para obtener r_c dado el valor de inicio r_o. Para todos estos cambios en la extensión de la prueba, la fórmula de Spearman-Brown supone que los reactivos añadidos (o eliminados en caso de acortar la prueba) son equivalentes a los otros reactivos de la prueba.

¡Inténtalo!

Practica con la fórmula de Spearman-Brown en el siguiente ejemplo: una prueba de 20 reactivos tiene

una confiabilidad original de consistencia interna de .75. Supón que la prueba aumenta al doble de extensión (es decir, n = 2). ¿Cuál es la confiabilidad de la prueba aumentada al doble?

Formulas de Kuder-Richardson

Una serie de fórmulas desarrolladas por G. Fredrick Kuder y M. W. Richardson (1973) proporcionan otras medidas de consistencia interna. Dos de estas fórmulas, 20 y 21, por lo general citadas como KR-20 y KR-21, se han usado mucho, por lo que las presentamos aquí. KR-20, la de mayor uso de las dos, se define así:

$$r_{KR\ -20} = \left(\frac{K}{K\ -\ 1}\right)\left(1\ -\ \frac{\sum pq}{DE_x^2}\right)$$

Fórmula 4-14

K = número de reactivos de la prueba
p = porcentaje de respuestas correctas
q = (1 – p)
DE_X = desviación estándar de las puntuaciones de la prueba

¿Qué es *pq*? En los reactivos que se califican de manera dicotómica, los que tienen respuestas del tipo correcto-incorrecto o sí-no, las calificaciones posibles son 1 o 0. *p* es el porcentaje de reactivos calificados con "1" –es decir, cuya respuesta es correcta o "sí"– mientras que *q* es simplemente (1 – *p*). Obtenemos *pq* para cada reactivo y luego sumamos estos valores de todos los reactivos de la prueba. El cuadro 4-9 presenta un ejemplo sencillo.

Cuadro 4-9. Datos muestra para determinar la confiabilidad de KR-20

Examinado	**Reactivos**					
	1	**2**	**3**	**4**	**5**	**Puntuación total**
A	1	1	1	1	1	5
B	1	1	1	1	0	4
C	1	0	1	0	0	2
D	1	1	0	0	0	2
E	1	1	1	1	1	5

He aquí una curiosa propiedad de KR-20. Recordemos la discusión acerca de dividir una prueba en mitades. Un método muy común es dividir la prueba en reactivos pares y reactivos nones; sin embargo, existen muchas otras divisiones posibles. Por ejemplo, una prueba de 10 reactivos podríamos dividirla en reactivos del 1 al 5 y del 6 al 10, o 1, 2, 5,

6, 9 en una mitad y 3, 4, 7, 8, 10 en la otra, o 1, 2, 3, 9, 10 y 4, 5, 6, 7, 8, y así sucesivamente. La fórmula KR-20 produce la correlación promedio entre todas las posibles mitades de la prueba.

Al empezar con la fórmula KR-20, asumimos que todas las "*p*" son iguales, es decir, todos los reactivos tienen el mismo porcentaje de respuestas "correctas" o de "sí". Recordemos que la suma de una constante (*C*) sobre *n* objetos es igual a $n \times C$. Por ejemplo, supongamos que la constante es 3; si sumamos la

constante sobre cinco objetos, obtenemos $3 + 3 + 3 + 3 + 3 = 5 \times 3$. Aplicando el principio a "*pq*" cuando todas las "*p*" son iguales, [Σ]*pq* se convierte en *npq*. Ya que $np = M$ (la media de las puntuaciones de la prueba), bajo el supuesto de que todas las "*p*" son iguales, la fórmula KR-20 puede escribirse como la KR-21:

$$r_{KR\,-\,21} = \left(\frac{K}{K\,-\,1}\right)\left(1\,-\,\frac{M(K\,-\,M)}{KDE_x^2}\right)$$

Fórmula 4-15

n = número de reactivos
M = media de las puntuaciones totales en la prueba
DE_x = desviación estándar de las puntuaciones de la prueba

El supuesto de que todas las "*p*" son iguales es bastante irreal. Si el supuesto se acercara a la verdad, el uso de KR-21 podría ser muy atractivo, porque es más fácil de calcularla que KR-20. La facilidad en los cálculos fue un criterio pertinente en la era previa a la computadora, pero ahora no es importante. De ahí que encontraremos confiabilidades KR-21 en manuales de pruebas y artículos de revista antiguos, pero no en los trabajos contemporáneos. No obstante, es útil recordar KR-21, pues permite estimar la confiabilidad cuando sólo *M* y DE_x están disponibles (por lo general, lo están) y no se puede obtener ninguna otra estimación de la confiabilidad. Thorndike (1982) señaló que KR-21 se aproxima mucho a KR-20 aun cuando las "*p*" varían mucho.

Coeficiente alpha

Las fórmulas de Kuder-Richardson requieren de reactivos que se califiquen de manera dicotómica. Existe una fórmula más general que no tiene esta restricción, pues los reactivos pueden tener cualquier tipo de calificación continua. Por ejemplo, los reactivos de una escala de actitud pueden calificarse con una escala de cinco puntos que va desde "totalmente en desacuerdo" (1) hasta "totalmente de acuerdo" (5). La forma más general es el **coeficiente alpha**, a menudo llamado **alpha de Cronbach** (véase Cronbach, 1951). Se debe tener cuidado de no confundir esta alpha con el alpha que usamos en las pruebas de significancia, porque no tienen nada que ver una con otra. Las dos versiones

equivalentes de la fórmula del coeficiente alpha son:

$$\alpha = \left(\frac{K}{K - 1}\right)\left(\frac{DE_X^2 - \sum DE_r^2}{DE_X^2}\right)$$

Fórmula 4-16

y

$$\alpha = \left(\frac{K}{K - 1}\right)\left(1 - \frac{\sum DE_r^2}{DE_X^2}\right)$$

K = número de reactivos de la prueba
DE_X = desviación estándar de las puntuaciones de la prueba
DE_r = desviación estándar de las puntuaciones de los reactivos

Podemos observar la semejanza de la notación entre estas fórmulas y la de KR-20. De hecho, cuando los reactivos se califican de manera dicotómica, [α] = $r_{KR\text{-}20}$, ya que para los reactivos que se califican de este modo (0, 1), $DE^2_r = pq$, así que [Σ]DE_r = [Σ]pq. Revisa un libro de estadística básica para verificar que la varianza de un determinado porcentaje es pq o $p(1 - p)$. El coeficiente alpha es muy usado en las pruebas contemporáneas actuales. Hogan, Benjamin y Brezinski (2000) encontraron que el coeficiente alpha se informa en más de dos terceras partes de las pruebas incluidas en el *Directoy of Unpublished Experimental Mental Measures*. Así, aunque el coeficiente alpha no es fácil de comprender, es importante para el estudiante de psicología estar familiarizado con él.

¿Qué indica el coeficiente alpha? Una forma alternativa de las que ya hemos presentado ayuda a responder esta pregunta. Suponiendo que todos los reactivos están "estandarizados", es decir, convertidos en una forma que tenga una media = 0 y una *DE* = 1, la siguiente fórmula se puede aplicar:

$$\alpha = \frac{K\,(\bar{r}_{ij})}{1 + (\,K\text{-}1\,)\bar{r}_{ij}}$$

Fórmula 4-17

donde

r_{ij} = correlación entre los reactivos i y j

K = número de reactivos

¿Qué hace esta fórmula? Sin duda no es evidente con sólo inspeccionar los elementos de la fórmula. Recordemos los fundamentos del método de división por mitades; era como crear miniformas alternas de la prueba. Ahora extendamos este razonamiento a los reactivos individuales; cada reactivo puede pensarse como una miniforma de la prueba. Entonces, podemos preguntar cómo cada una de estas miniformas (reactivos) concuerda con todas las demás miniformas de la prueba. Después, podemos sumar toda esta información en una medida de confiabilidad de consistencia interna. En esta fórmula, r_{ij} es la intercorrelación promedio entre todos los reactivos. Sin duda, no es una fórmula conveniente para propósitos de cálculo prácticos. Sin embargo, proporciona una mejor idea de lo que indican fórmulas anteriores KR-20 y el coeficiente alpha.

Las aplicaciones de la última fórmula ofrecen cierta orientación práctica acerca de cómo funciona la consistencia interna de las pruebas. Introduce los valores muestra para K y r_{ij} en la fórmula y observa las consecuencias. En el cuadro 4-10, asignemos a K los valores 5, 20 y 50, y a r_{ij}, los valores .10, .25 y .40. Entonces, calculamos [α]. ¿Qué podemos observar?

Primero, a medida que el número de reactivos aumenta, también aumenta la confiabilidad. Segundo, a medida que la correlación inter-reactivo aumenta, también aumenta la confiabilidad. Además, cuando hay relativamente pocos reactivos (5, por ejemplo), la confiabilidad es muy baja si las correlaciones inter-reactivo son bajas; cuando las correlaciones inter-reactivo son altas, la confiabilidad es mucho mayor, pero todavía no es muy alta. Cuando hay un gran número de reactivos (digamos, 50), la confiabilidad es muy respetable aun cuando las correlaciones inter-reactivo sean relativamente bajas. Así, la fórmula muestra que alpha depende de la correlación promedio entre los reactivos. El número de reactivos también es muy importante; alpha indica el grado en que los reactivos miden el mismo constructo o rasgo. A veces, esto se denomina medida de *homogeneidad de reactivos*, es decir, el grado en que los reactivos son iguales en términos de lo que miden. Podemos notar que los reactivos individuales no son muy confiables por sí mismos; de ahí que la intercorrelación entre ellos parezca, por lo general, baja. Por ejemplo, una correlación de .25 suele considerarse muy baja, pero es un nivel respetable si se trata de la correlación entre reactivos individuales.

En relación con las fuentes que atentan contra la confiabilidad que esbozamos anteriormente, el coeficiente alpha se relaciona con el muestreo del contenido. No mide la falta de confiabilidad debida a cambios en la aplicación de la prueba, condiciones personales o calificación. Esta misma generalización se puede aplicar a todos los métodos de consistencia interna para determinar la confiabilidad. He aquí la cuestión práctica: las medidas de confiabilidad de consistencia interna son fáciles de obtener y, por lo tanto, se citan mucho en los informes, a menudo como *la* confiabilidad de la prueba. Sin embargo, no dicen nada acerca de otras fuentes que afectan la confiabilidad, como inestabilidad temporal debida a fluctuaciones normales en las condiciones personales. Así que, cuando veamos, por decir algo, el coeficiente alpha, no debemos asumir que nos dice algo sobre

la estabilidad temporal.

¡Inténtalo!

Sustituye estos valores en la fórmula 4-17:
K = 30, r_{ij} = .10. ¿Cuál es el α?

Las diversas medidas de consistencia interna *no* son apropiadas para *pruebas de velocidad*; de hecho, son por completo inapropiadas si la prueba es primordialmente de velocidad, como las de velocidad de lectura o velocidad en tareas de oficina. Algunas pruebas de "poder" son, en parte, de velocidad, pues algunos examinados no terminan todos los reactivos. El grado en que la velocidad afecta la puntuación es también el grado en que las medidas de consistencia interna producirán estimaciones infladas de la confiabilidad. Para tratar este problema, es posible dividir las pruebas en términos de tiempo más que de número de reactivos, pero esto tiende a crear una situación de prueba un tanto artificial. Cuando la velocidad es un factor importante para determinar las puntuaciones, es mejor usar simplemente otros métodos de confiabilidad.

Tres conclusiones importantes

Como señalamos antes, no es fácil ver qué es exactamente lo que hacen las fórmulas de consistencia interna. Sin embargo, es fácil deducir tres conclusiones importantes al inspeccionar dichas fórmulas. Primero, la extensión de la prueba es importante; el número de reactivos siempre forma parte de las fórmulas. En general, mientras más extensa sea la prueba, más confiable será; las que son muy cortas a menudo no son confiables. En el caso de una prueba corta en extremo, los reactivos únicos casi siempre tienen una confiabilidad limitada. Como regla general, para aumentar la confiabilidad se debe aumentar la extensión de la prueba.

La segunda conclusión es que la confiabilidad se maximiza cuando se acerca a .50 el porcentaje de examinados que responde de manera correcta en una prueba de capacidades cognitivas o que responde en cierto sentido (p. ej., "sí") en una prueba no cognitiva. Podemos notar que pq alcanza su máximo cuando p = .50; pq disminuye conforme p se aleja de .50. Ésta es la razón de que las pruebas estandarizadas del campo de la cognición a menudo parezcan tan difíciles: el creador de la prueba trata de maximizar la confiabilidad. En realidad, tomando en cuenta el efecto de adivinar la respuesta correcta, el valor meta de p para los reactivos suele fijarse por arriba de .50, pero aún así a un nivel difícil. Sin embargo, Thorndike (1982) mostró que se sacrifica una pequeña parte de la confiabilidad al desviarse de manera considerable de p = .50. Retomaremos este tema en el capítulo 6.

Tercero, la correlación entre los reactivos es importante. Podemos observar el efecto de la correlación inter-reactivo promedio en el cuadro 4-10. La enseñanza práctica que

debemos recordar es ésta: para obtener una buena confiabilidad de consistencia interna, debemos usar reactivos que midan un rasgo bien definido.

Cuadro 4-10. Efecto del número de reactivos (K) y la correlación promedio inter-reactivos (r_{ij}) en el coeficiente alpha

K	r_{ij}	**α**
5	.10	.3 6
5	.25	.63
5	.40	.77
20	.10	.69
20	.25	.87
20	.40	.93
50	.10	.85
50	.25	.94
50	.40	.97

Error estándar de medición «93a

Un coeficiente de confiabilidad proporciona información valiosa sobre una prueba. Sin embargo, sus implicaciones prácticas para interpretar la prueba no son de inmediato evidentes. Para la interpretación práctica, dependemos del **error estándar de medición** (*EEM*), el cual se define así:

$$EEM = DE_x \sqrt{1 - r_{xx}}$$

Fórmula 4-18

donde r_{xx} es la confiabilidad de la prueba y DE_x, la desviación estándar en el grupo en que se determinó *r*.

El *EEM* es la desviación estándar de un número hipotéticamente infinito de puntuaciones obtenidas alrededor de la puntuación verdadera de una persona. Regresemos a la figura 4-8. Cada una de estas distribuciones tiene una desviación estándar. Este tipo de desviación estándar se denomina error estándar de medición. La distribución de la derecha en la figura 4-8 tiene un *EEM* relativamente grande. La distribución de la izquierda tiene un *EEM* relativamente pequeño. Podemos observar algunas de las consecuencias de la fórmula del *EEM*: si la confiabilidad de la prueba es perfecta ($r = 1.00$), el $EEM = 0$, es decir, no hay error de medición. ¿Cuál es el *EEM* si la confiabilidad de la prueba es .00, es decir, no es confiable en absoluto? En este caso, el *EEM* es la *DE* de la prueba.

¡Inténtalo!

Determina el EEM de los siguientes casos:
La confiabilidad de la ABC Test es .92; la DE es 15. ¿Cuál es el EEM?
Supón que la confiabilidad es .70, mientras que la DE se mantiene en 15. ¿Cuál es el EEM ahora?

Intervalos de confianza

El *EEM* puede emplearse para crear un intervalo de confianza, que en el lenguaje de las pruebas a veces se denomina **banda de confianza**, alrededor de la puntuación observada. Ya que el *EEM* es una desviación estándar de una distribución que, suponemos, es normal, se pueden aplicar todas las relaciones habituales. Regresemos a la curva normal de la figura 3-10b para refrescar la memoria. Por ejemplo, en 68% (cerca de dos terceras partes) de los casos, la puntuación verdadera estará dentro de +/−1 *EEM* de la puntuación observada. A la inversa, en cerca de una tercera parte de los casos, la puntuación observada diferirá de la puntuación verdadera por al menos 1 *EEM*.

Los informes sobre las puntuaciones generados por computadora a menudo utilizan la banda de confianza. El cuadro 4-11 muestra un ejemplo. La banda de confianza de la prueba A varía de 9 a 17, alrededor de la puntuación observada 13. La prueba B tiene una banda que va de 24 a 36, alrededor de la puntuación observada 30. Tales informes suelen citar la "banda" como +/−1 *EEM*, en esencia un intervalo de confianza de 68%, aunque también es fácil emplear una banda de 95% (+/−1.96 *EEM*) o de 99% (+/−2.58 *EEM*).

Cuadro 4-11. Muestra de un informe de puntuaciones con bandas de confianza

Prueba A	< < < <> > > >										
Prueba B					< < < < < <> > > > > >						
	0	5	10	15	20	25	30	35	40	45	50
				Puntuación de la prueba							

Unidades apropiadas para el EEM

El *EEM* debe expresarse en las unidades empleadas en la interpretación. Los manuales de las pruebas a menudo citan el *EEM* sólo en unidades de puntuación natural. Si en la interpretación se habla de puntuaciones normalizadas, la puntuación natural del *EEM* debe convertirse en normalizada. Esto se puede hacer con facilidad si las puntuaciones normalizadas son conversiones lineales de las naturales, como las puntuaciones estándar lineales. La tarea es mucho más complicada si se trata de conversiones no lineales. Por ejemplo, los rangos percentiles son conversiones no lineales a causa de la marcada desigualdad de las unidades percentiles (suponiendo una distribución más o menos

normal de las puntuaciones).

> El error estándar de medición... debe proporcionarse en unidades de la puntuación que se informa.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Consideremos los siguientes ejemplos de una prueba de 100 reactivos. La figura 4-9 muestra la distribución de las puntuaciones naturales de esta prueba (M = 75, DE = 5), las puntuaciones estándar (M = 500, DE = 100) y los percentiles. La confiabilidad de la prueba es .80; así que, en unidades de puntuación natural, el

$$EEM = 5\sqrt{1-.80} = 2.2.$$

Esta puntuación es igual a 44 unidades en el sistema de puntuaciones estándar, es decir,

$$100\sqrt{1-.80}$$

Es evidente que el *EEM* de la puntuación natural no es útil si la interpretación se basa en las puntuaciones estándar. No hay una conversión sencilla para los rangos percentiles; sin embargo, podemos estimar el efecto de aplicar +/−1 *EEM* en las unidades de puntuación natural en varios puntos a lo largo de la escala de percentiles. Alrededor del percentil 5 (o 95), +/−2.2 unidades de puntuación natural cubren casi 10 puntos percentiles, mientras que alrededor del percentil 50, ¡+/−2.2 de estas unidades cubren 34 puntos percentiles! Esta diferencia surge por la misma razón que discutimos en el capítulo 3 en relación con la interpretación de los percentiles, pero aquí se aplica a la interpretación del *EEM*.

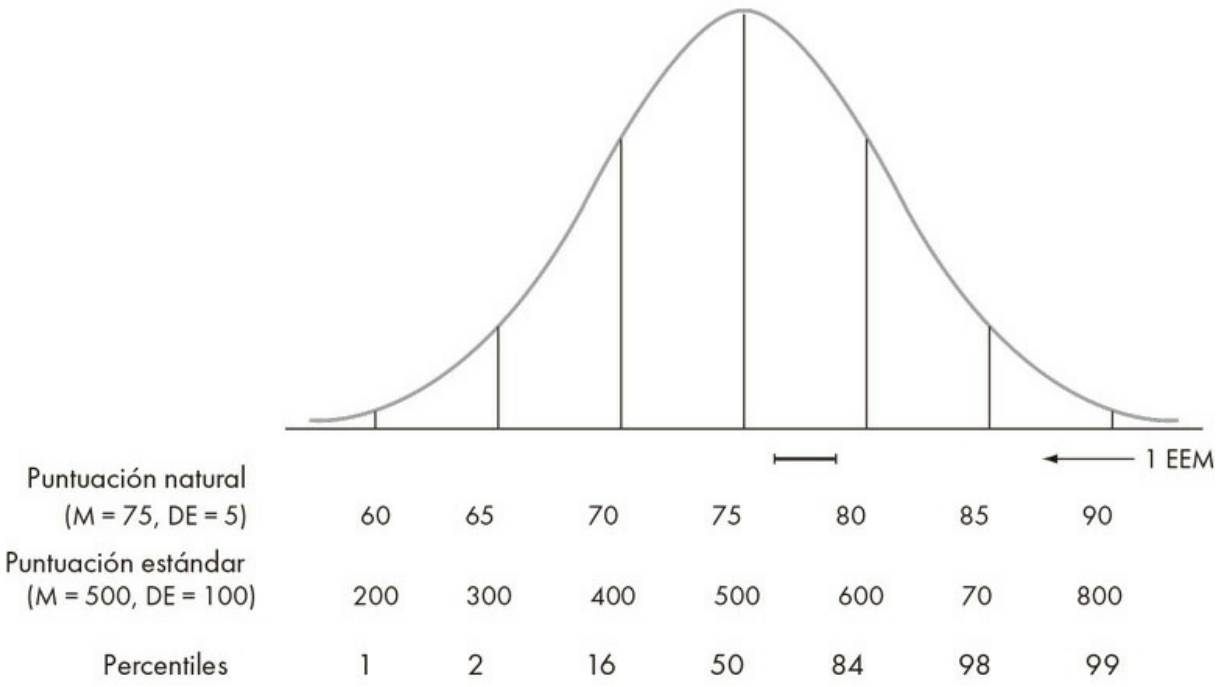


Figura 4-9. *EEM* en diferentes unidades de puntuación.

Error estándar de la diferencia

En la sección anterior, describimos el error estándar de medición para una puntuación única. ¿Qué pasa cuando se comparan dos puntuaciones? ¿Cómo debe aplicarse el concepto de error de medición en tal caso? ¿Se suman los errores estándar de puntuaciones separadas? ¿Se toma su promedio? Para responder a estas preguntas, las cuales no son evidentes para el sentido común, presentamos la siguiente fórmula:

$$EEM_{dif} = \sqrt{EEM_1^2 + EEM_2^2}$$

Fórmula 4-19

donde

EEM_{dif} = **error estándar de la diferencia** entre dos puntuaciones
EEM_1 = error estándar de la primera prueba
EEM_2 = error estándar de la segunda prueba

Recordemos que

$$EEM_1 = DE_1\sqrt{1-r_{11}} \text{ y } EEM_2 = DE_2\sqrt{1-r_{22}}$$

A menudo sucede que $r_{11} = r_{22}$ y $DE_1 = DE_2$. Si ése es el caso, la fórmula de EEM_{dif} se simplifica de la siguiente manera:

$$\text{EEM}_{dif} = \text{DE}\sqrt{2(1 - r)}$$

Fórmula 4-20

donde

DE = desviación estándar común
r = coeficiente de confiabilidad común

Suponemos que la distribución de diferencias entre puntuaciones es normal y que EEM_{dif} es la desviación estándar de esta distribución. De ahí que todas las afirmaciones habituales acerca de la desviación estándar se puedan aplicar aquí: 68% de los casos caen dentro de +/− 1 *DE*, 5% cae fuera de +/−1.96 *DE*, y así sucesivamente.

Tres tipos de errores estándar

El *error estándar de medición* debe distinguirse con cuidado de otros dos tipos de errores estándar: el *error estándar de la media* y el *error estándar de estimación*. Estas distinciones son fuente de gran confusión para los psicómetras novatos, sobre todo porque cada una de estas entidades puede nombrarse de manera abreviada sólo como "error estándar" y se espera que sepamos, con base en el contexto, cuál de ellas puede aplicarse. Las tres son, en verdad, desviaciones estándar, pero lo son de diferentes cosas. Bosquejaremos brevemente las diferencias entre estos tres tipos de errores estándar.

Resumen de puntos clave 4-5

Tres tipos de errores estándar
1. Error estándar de medición:
Índice del error debido a la falta de confiabilidad (fórmula 4-18)
2. Error estándar de la media:
Índice del error debido al muestreo aleatorio (fórmula 4-21)
3. Error estándar de estimación:
Índice del error en la predicción de Y a partir de X (fórmula 4-4)

El **error estándar de medición** es la desviación estándar de una población hipotética de puntuaciones observadas distribuidas alrededor de la puntuación verdadera de un

individuo. En la figura 4-8 presentamos ejemplos de estas distribuciones. La fórmula pertinente es la 4-18. El **error estándar de la media** es la desviación estándar de una población hipotética de medias muestrales que corresponde a muestras (de cierto tamaño) distribuidas alrededor de la media poblacional. El error estándar de la media se usa en las pruebas de significancia estadística, por ejemplo, prueba *t*, prueba *z*, y para los intervalos de confianza de las medias muestrales. Recordemos de la estadística básica que el error estándar de la media es:

$$EE_X = \frac{DE_X}{\sqrt{N}}$$

Fórmula 4-21

donde

DE_X = desviación estándar de las puntuaciones
N = tamaño de la muestra

El **error estándar de estimación** (a veces también llamado error estándar de predicción) es la desviación estándar de las puntuaciones *Y* reales alrededor de las puntuaciones *Y* predichas cuando *Y* se predice a partir de *X*. Encontramos el error estándar de estimación en nuestro repaso de estadística antes en este capítulo. Su fórmula es la 4-4.

Es importante tener en mente estas distinciones. Las diferencias entre estos tres tipos de errores estándar tienen consecuencias reales en la práctica. No se trata de una sutileza académica ni de ser quisquillosos por el puro placer de serlo.

Algunos temas especiales relacionados con la confiabilidad «96a

Confiabilidad en los informes interpretativos

La información de la confiabilidad suele aparecer en términos cuantitativos precisos, es decir, en forma de coeficientes de confiabilidad y errores estándar de medición. Sin embargo, cada vez más el desempeño en la prueba se informa con una narración, a menudo llamada informe interpretativo, el cual puede aligerar mucho la tarea de interpretar las puntuaciones de la prueba. Desafortunadamente, los informes narrativos no se adaptan con facilidad a las herramientas tradicionales del análisis de confiabilidad. Algunos informes narrativos incorporan con claridad los conceptos de confiabilidad y errores de medición, pero otros no. Los informes pueden dar la impresión de que la confiabilidad no es importante, aunque en realidad siempre lo es. El lector del informe narrativo debe asegurarse de que a) conoce la información de confiabilidad acerca de la

prueba y b) utiliza la información cuando interpreta el informe. Incluso los informes narrativos deben incorporar el concepto de error de medición.

Confiabilidad de subpuntuaciones y reactivos individuales

Debe proporcionarse la información de confiabilidad de la "puntuación" que, en realidad, se está interpretando. Consideremos el siguiente ejemplo. Una batería tiene cuatro pruebas separadas; el manual de la prueba ofrece información de confiabilidad de cada prueba; todas tienen una buena confiabilidad, digamos $r > .90$. Sin embargo, los informes de puntuaciones de la batería pueden dar información sobre el desempeño del individuo en grupos de reactivos o, incluso, en reactivos individuales de las pruebas. No podemos asumir que estos grupos o reactivos individuales tienen la misma confiabilidad que las puntuaciones totales de las pruebas. De hecho, es muy cierto que el desempeño en los grupos o reactivos individuales será, por mucho, menos confiable que las puntuaciones totales. La confiabilidad de los grupos de, digamos, tres o cuatro reactivos es notablemente baja, en el mejor de los casos alrededor de .30 a .40. Por lo general, no consideraríamos usar una prueba con una confiabilidad de .30. Lamentablemente, los grupos de reactivos con confiabilidades en este rango aparecen de manera cotidiana en los informes. El desempeño en los reactivos individuales es aun menos confiable. ¡Hay que ser cuidadosos con esto!

Confiabilidad de los perfiles

Los perfiles de las puntuaciones a menudo son la base para interpretar las pruebas. En la figura 4-10 aparecen perfiles de muestra de una batería de cuatro pruebas. Lo que puede ser interesante aquí no es el nivel absoluto de las puntuaciones de las pruebas, sino los *patrones* que se despliegan en los perfiles. Por ejemplo, la "V" formada por las puntuaciones en las pruebas A-C de Sue y Fred puede ser de especial interés. La confiabilidad de tales patrones no se puede representar con facilidad, pero, sin duda, es menos confiable que las pruebas individuales. Este tema se relaciona con el error estándar de la diferencia tratado antes; señalamos que el error de medición en el caso de las diferencias combina los errores en las puntuaciones individuales. Esta composición de la falta de confiabilidad es aún mayor cuando un perfil de tres o más puntuaciones es la base de la interpretación.

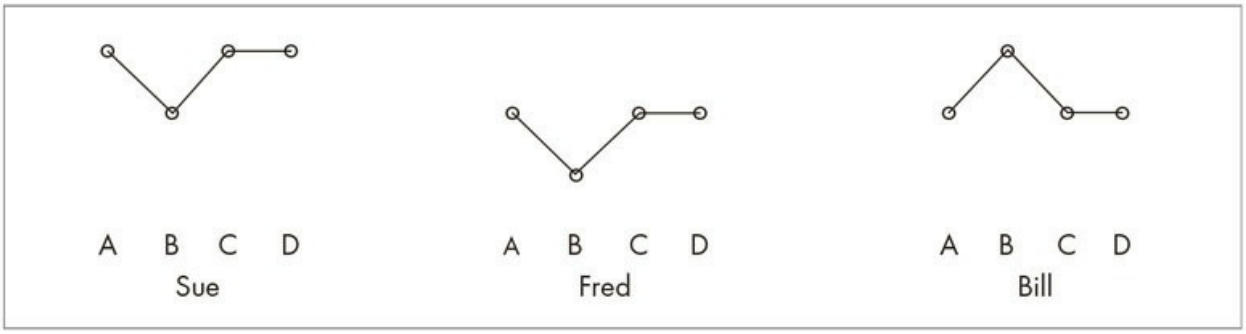


Figura 4-10. Perfiles muestra de las puntuaciones de las pruebas A, B, C y D.

Confiabilidad de las pruebas con referencia a un criterio

Recordemos la distinción entre pruebas referidas a un criterio (PRC) y pruebas referidas a una norma (PRN), la cual vimos en el capítulo 3. La diferencia clave está en el método de interpretación. Los métodos para determinar la confiabilidad pueden no ser diferentes para las PRC, dependiendo de la distribución de las puntuaciones de la prueba y de los usos de dichas puntuaciones. Los enfoques clásicos de la confiabilidad revisados en este capítulo suponen una distribución normal de las puntuaciones. Al menos, debe haber una distribución razonable de las puntuaciones; de lo contrario, el coeficiente de confiabilidad no funciona. Consideremos el caso extremo en un estudio de test-retest donde todas las puntuaciones son las mismas, digamos que son perfectas, en la segunda aplicación. La fórmula del coeficiente de correlación nos llevará a $r = .00$.

La preocupación por la variabilidad inadecuada en las puntuaciones de la prueba *puede* ser aplicable a algunas situaciones de dominio de las pruebas donde las distribuciones de puntuaciones tienen una marcada asimetría negativa, es decir, una acumulación de puntuaciones en la puntuación perfecta o cerca de ella. La distribución de las puntuaciones de PRC en relación con un "punto de corte" también puede afectar la interpretación de una PRC. Se han desarrollado numerosos métodos para expresar la confiabilidad de las PRC en estas circunstancias especiales; se pueden consultar en Berk (1984), Crocker y Algina (1986), Feldt y Brennan (1989) y Nunnally y Bernstein (1994). No encontramos con frecuencia estos métodos especializados en la práctica.

Confiabilidad en la teoría de la respuesta al reactivo «97a

La confiabilidad es todo un tema en las pruebas construidas de acuerdo con la teoría de la respuesta al reactivo (TRR), del mismo modo que en las pruebas elaboradas de acuerdo con la teoría clásica de las pruebas (TCP). Nuestro tratamiento de la confiabilidad se ha concentrado en el enfoque de la TCP, porque la gran mayoría de las pruebas existentes sigue este enfoque. Sin embargo, cada vez más pruebas siguen el enfoque de la TRR; de ahí que debamos examinar la confiabilidad en el contexto de la TRR. Primero, notemos que el enfoque único de la TRR en relación con la confiabilidad se ocupa sólo de la consistencia interna. Aun cuando una prueba sea construida y calificada de acuerdo con la TRR, si la preocupación gira en torno a la estabilidad temporal o a la consistencia entre los jueces, deben usarse los métodos descritos antes en este capítulo para determinar la confiabilidad.

Para los propósitos del análisis de la consistencia interna, la TRR proporciona un enfoque diferente de los que ya hemos descrito. Como el coeficiente alpha, el análisis de la confiabilidad en la TRR depende del funcionamiento de los reactivos dentro de la prueba. Sin embargo, en la TRR los reactivos funcionan de manera independiente, mientras que en el análisis de la consistencia interna de la TCP, los reactivos son interdependientes.

En la TRR, el error estándar se expresa como:

$$EE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Fórmula 4-22

donde θ es la puntuación (theta) de la capacidad o rasgo, descritos en el capítulo 3, e I(θ) es la función de información de la prueba, la cual es simplemente la suma de las funciones de la información de los reactivos; estas últimas surgen de las características de los reactivos, las cuales se describen con mayor detalle en el capítulo 6.

El error estándar *EE*(θ) en la TRR se menciona a menudo como un índice de la **precisión de la medición**. Tiene una ventaja importante sobre el *EEM* en la TCP, donde se supone que el *EEM* es el mismo en todos los niveles de puntuación.[4] Consideremos una prueba de CI con *DE* = 15 y [α] = .89, de modo que *EEM* = 5. Este *EEM* se aplica en todo el rango del CI, 80, 100, 150, etc. Además, depende no sólo de la homogeneidad de los reactivos de la prueba, sino también de la heterogeneidad de los individuos con los que se determinó [α]. *EE(θ)* no tiene estas limitaciones, pues se determina para cada nivel específico de las puntuaciones, es decir, para cada nivel de theta. Así, para una

prueba particular, *EE(θ)* puede ser relativamente menor en el caso de las puntuaciones bajas y relativamente mayor en el caso de las altas, o viceversa, dependiendo de cómo funcionen los reactivos en varios niveles del rasgo. En el capítulo 9 (250a»), presentamos un ejemplo de esta aplicación con el GRE-General Test, en el cual el error es mayor en las puntuaciones de nivel medio que en las puntuaciones extremas. En otras pruebas, *EE(θ)* puede ser menor en las puntuaciones de nivel medio. En Hambleton y Swaminathan (1985), Hambleton, Swaminathan y Rogers (1991) y deAyala (2009) se pueden encontrar más detalles en relación con la confiabilidad en la TRR, mientras que en Thissen (2000) se puede consultar un tratamiento detallado de la confiabilidad de pruebas de aplicación adaptable por computadora basadas en la TRR.

Resumen de puntos clave 4-6

Temas especiales relacionados con la confiabilidad

- Informes interpretativos
- Subpuntuaciones y reactivos individuales
- Perfiles
- Pruebas referidas a un criterio

Teoría de la generalizabilidad

De la revisión de los distintos tipos de confiabilidad que presentamos en este capítulo, debería ser claro que no existe algo que pueda considerarse *la* confiabilidad de una prueba. Hay muchas fuentes que atentan contra la confiabilidad. Cada método para determinar la confiabilidad intenta tratar una o unas pocas de estas fuentes; de ahí que podamos decir que una prueba tiene una confiabilidad de test-retest de .85, una confiabilidad de formas alternas de .78 y un coeficiente alpha de .92. Cada una de ellas puede determinarse en un estudio separado. La **teoría de la generalizabilidad** (TG) intenta evaluar varias fuentes de falta de confiabilidad al mismo tiempo.

La TG empieza con la misma noción básica que la teoría clásica de las pruebas, es decir, que cada persona tiene una puntuación verdadera, la cual, en la TG, se denomina a menudo **puntuación de universo** o puntuación de dominio. Pensemos en una persona que es evaluada en varias ocasiones con muchas formas y puntuaciones diferentes. La puntuación verdadera de la persona o puntuación de universo es la puntuación promedio de todas las ocasiones en que se aplicaron las pruebas. Ahora imaginemos 500 personas respondiendo estas evaluaciones múltiples. En el caso de cualquier par específico de evaluaciones, podríamos determinar una de las medidas clásicas de confiabilidad; por ejemplo, la correlación entre puntuaciones de dos ocasiones sería la confiabilidad de test-retest. Sin embargo, sería muy útil si pudiéramos determinar, *en un solo estudio*, la confiabilidad para varias ocasiones, varias formas y varios jueces. Esto es lo que la TG intenta hacer.

El análisis de varianza (ANOVA) brinda el marco básico para el estudio de generalizabilidad (estudio G). Recordemos que el ANOVA permite estudiar el efecto de diversas variables independientes de manera simultánea sobre una variable dependiente y las interacciones, es decir el efecto único creado por la combinación de dos (o más) variables independientes muy aparte de sus efectos por separado.

> ... la teoría de la generalizabilidad impulsa al investigador a especificar y estimar los componentes de la varianza de la puntuación verdadera, la varianza de la puntuación de error y la varianza de la puntuación observada, y a calcular los coeficientes con base en estas estimaciones, las cuales suelen realizarse aplicando técnicas del análisis de varianza.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Supongamos que estamos midiendo ansiedad. Tenemos una muestra de 50 personas y las examinamos en cinco ocasiones diferentes. En cada una, les presentamos dos tareas que podrían aumentar su ansiedad. Tenemos cuatro jueces valorando el grado de ansiedad manifiesta. Este diseño nos permite investigar la consistencia:

- A lo largo de las cinco ocasiones (como en la confiabilidad de test-retest)
- Entre las tareas (como en la confiabilidad de formas alternas)
- Entre los jueces (como en la confiabilidad interjueces)

Esto da origen a un diseño factorial ANOVA de 5 × 2 × 4: ocasiones × tareas × jueces. Con este diseño, podemos estudiar la varianza debida a cada factor, así como a las interacciones entre los factores.

La literatura de la TG distingue entre un estudio de generalizabilidad (estudio G) y un estudio de decisión (estudio D). El estudio G analiza los componentes de la varianza, incluyendo las interacciones, mientras que el estudio D utiliza los resultados del estudio G para *decidir* cómo podría mejorarse la medición haciendo cambios en uno de los componentes. Pensemos en nuestro estudio sobre la ansiedad. Usamos cuatro jueces; ¿eso ofrece suficiente estabilidad en los resultados? ¿Tendríamos suficiente estabilidad si usáramos sólo dos jueces? Las respuestas a estas preguntas pueden ayudar a refinar y mejorar el proceso de medición.

Los detalles para llevar a cabo un análisis de generalizabilidad están fuera del alcance de este libro, pero el lector interesado en este tema puede consultar Shavelson, Webb y Rowley (1989) para tener un buen panorama de él, Shavelson y Webb (1991) para revisar un tratamiento más detallado y Brennan (2001b) para leer una exposición completa. Brennan (2001a, 2011) ofrece un análisis histórico que muestra el desarrollo de la TG desde los primeros métodos de análisis de confiabilidad. Brennan (2000) también advierte que el marco del ANOVA anteriormente descrito para la TG puede ser tomado con demasiada literalidad. No obstante, este marco constituye una buena introducción a la TG.

La teoría de la generalizabilidad ofrece un marco de excepcional utilidad para pensar en la confiabilidad de las medidas. Sin embargo, hasta ahora, no se ha empleado mucho en aplicaciones prácticas, probablemente porque es un fastidio llevar a cabo incluso los estudios sencillos de confiabilidad (excepto los de consistencia interna). ¿Quién quiere hacer la misma prueba en dos ocasiones diferentes? ¿O responder dos formas distintas de la misma prueba? Llevar a cabo un estudio que varía en, digamos, tiempo de aplicación, número de formas y procedimientos de calificación se vuelve muy difícil desde un punto de vista práctico. Sí encontramos estudios de generalizabilidad de unas pocas pruebas y, quizá, podamos ver, al menos, cierto aumento en el uso de esta metodología en el futuro. Schmidt, Le e Ilies (2003) sugirieron un intento en cierta forma más práctico de establecer múltiples fuentes que atentan contra la confiabilidad, pero ese método no ha conseguido mucha atención. Por el momento, el punto más importante es la perspectiva que la metodología nos ofrece acerca del campo entero de la confiabilidad.

Factores que afectan los coeficientes de confiabilidad

Recordemos nuestra discusión en este capítulo sobre los cuatro factores que afectan los coeficientes de correlación. Ya que la confiabilidad suele expresarse como un coeficiente de correlación, estos factores pueden afectar los datos de confiabilidad. Consideremos cada uno de estos factores.

Primero, el hecho de que la correlación sea una cuestión de posición relativa más que de puntuaciones absolutas *no* es una preocupación importante para la confiabilidad. Segundo, la curvilinealidad no es, por lo general, un tema para los datos de confiabilidad. Aunque en teoría es posible tener una tendencia curvilineal en los datos de confiabilidad, esto no suele ocurrir en la práctica. En cualquier caso, es fácil verificar el supuesto de linealidad examinando una distribución bivariada de los datos de confiabilidad. Cualquier paquete estadístico estándar desarrolla una distribución bivariada para inspeccionarla.

Tercero, la heterocedasticidad bien puede ser un problema para el error estándar de medición. Otra vez, la solución es realizar una gráfica bivariada y verificar el supuesto de homocedasticidad. Debemos hacer notar que la precisión de la medición estadística en la TRR, $EE(\theta)$, proporciona errores estándar diferentes de los distintos niveles de puntuación, de modo que se ajustan para cualquier falta de homocedasticidad.

Por último, la variabilidad grupal es *a menudo* un problema cuando se interpretan datos de confiabilidad, pues éstos se han desarrollado para un grupo mucho más homogéneo y más heterogéneo que el grupo considerado pertinente para nuestro marco interpretativo. La solución para este problema es usar las fórmulas 4-5 y 4-6 para corregir la homogeneidad o heterogeneidad excesiva. Esta corrección con frecuencia se usa en el trabajo práctico. En el capítulo 9, mostraremos dichas correcciones al usar pruebas para predecir el éxito escolar, académico y laboral.

¿Qué tan alta debe ser la confiabilidad?

Después, incluso, de la más breve exposición al tema de la confiabilidad, nos sentimos inclinados a preguntar: ¿qué tan alta debe ser la confiabilidad de una prueba? No hay respuesta más sencilla a esta pregunta que: depende. En particular, depende de qué queremos hacer con la prueba. Es como preguntar "¿qué tan alta debe ser una escalera?" Depende. ¿Es para cambiar un foco que no alcanzas o necesitas subir al techo de un edificio de tres pisos?

Si necesitamos tomar una decisión muy importante en la que la información sobre la prueba tenga mucho peso –por ejemplo, otorgar una licencia de ejercicio profesional en algún campo–, vamos a requerir de una prueba con una confiabilidad alta. Si la prueba es sólo una de muchas fuentes de información que nos darán una idea aproximada acerca del nivel general de ajuste de una persona, entonces un grado moderado de confiabilidad puede ser suficiente. Si la prueba se usa en un proyecto de investigación en el que los promedios grupales son el centro de atención, entonces un grado de confiabilidad aun menor será suficiente.

Todo mundo está de acuerdo con las generalizaciones que acabamos de citar; sin embargo, aún es útil tener en mente algunos puntos de referencia de la confiabilidad, los cuales podemos encontrar en numerosas fuentes (véase Charter, 2003; Groth-Marnat, 2009; Hunsley & Mash, 2008; Kaplan & Saccuzzo, 2013; Murphy & Davidshofer, 2001; Nunnally & Bernstein, 1994). Aquí presentamos nuestro resumen de lo que parece ser un consenso respecto al tema. Un coeficiente de confiabilidad de al menos .90 es excelente; se requiere este nivel o incluso .95 cuando la prueba tiene un gran peso para tomar una decisión importante, como ubicación en cursos, exámenes para autorizar el ejercicio profesional o la clasificación de una persona como intelectualmente discapacitada en un caso forense. La confiabilidad de .80 a .89 es buena; cuando una prueba tiene una confiabilidad en este rango, debe tomarse en cuenta otro tipo de información. Suponiendo que la otra información tiene una confiabilidad respetable, la combinación de ambas fuentes tiene una confiabilidad mayor. La confiabilidad de .70 a .79 es adecuada, pero el uso de la puntuación de la prueba requiere mucho cuidado y, sin duda, debe complementarse con información de otras fuentes. El uso de pruebas con confiabilidades en el rango de .60 a .69 deberá limitarse, quizá, a la investigación. Si la confiabilidad está por debajo de .60, deberíamos buscar otra prueba con una mejor confiabilidad. Sin embargo, podemos notar el principio general de que reunir diversas fuentes con confiabilidad limitada produce una combinación con mayor confiabilidad, una especie de versión generalizada de la fórmula de Spearman-Brown.

Aquí presentamos cinco puntos importantes que complementan la discusión sobre qué tan alta debe ser la confiabilidad. Primero, la mayoría de los informes de confiablidad cubren sólo una fuente (p. ej., consistencia interna o test-retest), pero en nuestra práctica debemos tomar en cuenta múltiples factores que influyen en la confiabilidad (como en el análisis de la teoría de la generalizabilidad, que casi nunca está disponible). Así, cuando

encontramos un coeficiente alpha de, digamos, .90, nos engañaríamos si pensamos que "tenemos todo resuelto" con respecto a la confiabilidad. Segundo, muchos usos de las pruebas tienen que ver con las diferencias entre puntuaciones, sea de manera directa o en la forma del perfil. Estas diferencias casi siempre son menos confiables que las confiabilidades de las pruebas que forman parte de las diferencias o los perfiles.

Tercero, a veces nos encontramos con el argumento de que la confiabilidad no es un tema importante para cierto tipo de pruebas o para cierta puntuación particular de una prueba. ¡Nunca debemos creer en tal afirmación! La confiabilidad siempre es importante. La información que no es confiable o cuya confiabilidad es desconocida no debe utilizarse. Cuarto, recordemos nuestra discusión de la relación entre la extensión de la prueba y la confiabilidad: las pruebas cortas suelen ser más bien no confiables. A veces nos encontramos con que el autor de una prueba o incluso de una reseña dice que la prueba tiene una confiabilidad bastante buena tomando en cuenta lo breve que es. Debemos tener cuidado con esta afirmación. Una prueba con una confiabilidad de .60 –breve o extensa– es una prueba con una confiabilidad de .60, la cual no es muy buena. Si esta prueba se tuviera que usar para propósitos serios, su confiabilidad debería aumentarse, quizá, haciéndola más extensa. Quinto, algunos autores de pruebas informan la significancia estadística de los coeficientes de confiabilidad señalando, a menudo con malicia, que el coeficiente es sumamente significativo. Tales informes no son muy útiles, pues tenemos estándares más elevados para los coeficientes de confiabilidad que la sola significancia estadística.

Más importante que la confiabilidad de la prueba es su validez. Aunque una prueba sin confiabilidad no puede tener validez alguna, es posible tener pruebas muy confiables que no sean válidas para los propósitos que tenemos en mente. Además, una prueba con una confiabilidad y validez moderadas es preferible que una prueba con una confiabilidad alta y una validez baja. Estas breves observaciones constituyen una transición hacia el tema crucial del siguiente capítulo: la validez de las pruebas.

Resumen

1. La confiabilidad, uno de los conceptos más importantes en el campo de las pruebas, se ocupa de la consistencia o replicabilidad de las puntuaciones de las pruebas.
2. Distinguimos entre confiabilidad y validez, el sentido psicométrico de la confiabilidad y diversos usos cotidianos del término, cambios reales y fluctuaciones temporales en las medidas, y errores constantes y errores no sistemáticos.
3. El coeficiente de correlación (*r*) es el método más común para expresar la confiabilidad; de ahí la importancia de comprender las correlaciones y los factores que influyen en ellas.
4. Las principales fuentes de varianza que afectan la confiabilidad son la calificación de la prueba, su contenido, las condiciones de aplicación y las condiciones personales del examinado.
5. La teoría clásica de las pruebas utiliza los conceptos de puntuación verdadera, puntuación de error y puntuación observada.
6. Entre los métodos de uso común para determinar la confiabilidad están el de test-retest, formas alternas, interjueces y varias medidas de consistencia interna. Cada método se ocupa de una o algunas fuentes que atentan contra la confiabilidad, pero no de todas.
7. El error estándar de medición (*EEM*) y los intervalos de confianza ayudan a traducir los coeficientes de confiabilidad en interpretaciones prácticas.
8. El concepto de error estándar se aplica no sólo a la interpretación de puntuaciones únicas, sino también a las diferencias entre puntuaciones y perfiles de puntuaciones.
9. El error estándar de medición debe distinguirse del error estándar de la media y del error estándar de estimación.
10. Los conceptos de confiabilidad y error estándar se aplican igualmente a los informes interpretativos y cuantitativos del desempeño en las pruebas.
11. La confiabilidad es importante para la interpretación con referencia a un criterio, pero la situación a veces requiere modificar el método usual para determinar la confiabilidad.
12. La teoría de la respuesta al reactivo (TRR) emplea el concepto de precisión de la medición, el cual puede diferir en varios puntos a lo largo de la escala.
13. Usando técnicas de análisis de varianza, la teoría de la generalizabilidad intenta abordar las diversas fuentes de falta de confiabilidad en un solo estudio.
14. Los factores que afectan los coeficientes de correlación, en especial la variabilidad grupal, deben tomarse en cuenta al interpretar los datos de la confiabilidad.
15. El uso que se le va a dar a la prueba determina el nivel de confiabilidad que se requiere. Para tomar decisiones importantes, la confiabilidad debe ser de al menos .90. En casos en que la prueba es una de varias fuentes de información que se consideran en conjunto, la confiabilidad deseada es de al menos .80.

Palabras clave

alpha de Cronbach
cambio real
coeficiente alpha
coeficiente de correlación
coeficiente de correlación intraclase
confiabilidad
confiabilidad de división por mitades
confiabilidad de formas alternas
confiabilidad de pares y nones
confiabilidad interjueces
confiabilidad test-retest
consistencia interna
corrección de Spearman-Brown
dispersograma
distribución bivariada
error constante
error estándar de estimación
error estándar de la diferencia
error estándar de la media
error estándar de medición
error no sistemático
heterocedasticidad
heterogeneidad
homocedasticidad
homogeneidad
intervalos de confianza
KR-20
KR-21
línea de regresión
precisión de la medición
puntuación de error
puntuación de universo
puntuación observada
puntuación verdadera
teoría de la generalizabilidad

Ejercicios

1. Usa algún programa de cómputo con el que estés familiarizado (p. ej., SPSS, Minitab, SAS o Excel) para obtener la correlación de estos datos.

Examinado	Prueba X	Prueba Y
1	20	24
2	18	12
3	23	27
4	34	37
5	19	15
6	33	45
7	16	10
8	35	42
9	15	10
10	22	24

2. Prepara una distribución bivariada para las puntuaciones del problema anterior.

3. Utilizando cualquier base de datos de tu biblioteca, haz una búsqueda por palabras clave introduciendo *test reliability* [confiabilidad de la prueba]. ¿Qué clase de referencias encontraste? (*Nota*: es probable que encuentres referencias de otros campos aparte de los relacionados con las pruebas psicológicas).

4. Usa la fórmula de Spearman-Brown (90a») con los siguientes ejemplos. Una prueba de 20 reactivos tiene una confiabilidad de consistencia interna original de .75.

a. ¿Cuál es r_c si la prueba cuadruplica su número de reactivos (80 reactivos, $n = 4$)?

b. ¿Cuál es r_c si la prueba reduce a la mitad su número de reactivos (10 reactivos, $n = .5$)?

c. Quieres que r_c sea de .90. ¿Cuántos reactivos debería tener la prueba? (Encuentra el valor de n, luego multiplica n por 20, la extensión original de la prueba).

5. Calcula $r_{KR\text{-}20}$ con los siguientes datos. Los números del cuadro indican respuestas correctas (1) e incorrectas (0). Algunos de los cálculos ya están hechos.

Reactivo	1	2	3	4	5	Puntuación total
Examinado						
A	1	1	1	1	1	5
B	1	1	1	1	0	4
C	1	0	1	0	0	2
D	1	1	0	0	0	2
E	1	1	1	1	1	5

F	1	0	1	1	1	4
G	1	1	1	0	1	4
H	0	0	0	1	0	1
I	0	1	0	0	0	1
J	0	0	0	0	0	0
p =	.7	.6	.6	.5	.4	M = 28 / 10 = 2.8 P_x = 1.81 (usando $n - 1$ en el denominador)
$\sum pq = (.7x.3) + (.6x.4) + (.6x.4) + (.5x.5) + (.4x.6) = 1.18$						

6. Con base en los datos del cuadro 4-9, ¿cuál es la p del reactivo 2? ¿Cuál es pq?
p = _____ pq = _____

7. Calcula el coeficiente alpha de los siguientes datos. Los números son las respuestas a reactivos de actitud, cada uno calificado en una escala de 5 puntos.

Reactivo	**1**	**2**	**3**	**4**	**5**	**Puntuación total**
Examinado						
A	5	4	5	3	5	22
B	4	4	3	4	4	19
C	4	3	3	4	4	18
D	3	3	3	4	3	16
E	3	3	3	3	3	15
F	3	3	3	2	2	13
G	2	2	2	2	2	10
H	2	1	2	1	2	8
I	1	2	2	1	1	7
J	1	1	2	1	1	6
Pr =	_	_	_	_	_	Px =

8. Usando la fórmula 4-17 de [α], completa el siguiente cuadro. (Esto requerirá de un poco de álgebra sencilla.)

K	$\mathbf{r}_{ij}$	**α**
10	.15	_
25	_	.90
_	.20	.80

9. Consideremos estos datos de una prueba: confiabilidad = .90, DE = 15.
a. ¿Cuál es el error estándar de medición (*EEM*)?
b. ¿Cuál es la probabilidad de que la puntuación verdadera de una persona se

encuentre dentro de +/−1 *EEM* con respecto a la puntuación obtenida?

c. ¿Cuál es el intervalo de confianza de 95% para estos datos?

10. Supón que dos pruebas tienen la misma *DE* = 10 y la misma confiabilidad (*r*) = .80. ¿Cuál es el EEM_{dif} de estas pruebas? (Usa la fórmula 4-20). Supón una *DE* = 10 y una *r* = .60 en común. Ahora, ¿cuál es el EEM_{dif} de estas pruebas?

11. Usa los datos del apéndice D2 para determinar la confiabilidad de test-retest de las medidas. Simplemente obtén las correlaciones entre la primera aplicación y la segunda utilizando tu programa de cómputo de estadística. ¿Qué concluyes acerca de estas confiabilidades?

Notas

1 El término **línea de regresión** no es muy descriptivo; sería más agradable llamarla "línea de predicción". Sin embargo, el término regresión se adoptó en etapas tempranas del desarrollo de esta metodología, lo que se originó con el trabajo de Francis Galton. El término ha tenido un poder impresionante, pero desafortunado, para sobrevivir.

2 Ya que el error (E) puede ser positivo o negativo, algunas fuentes escriben la fórmula 4-7 como T = O + E, con los cambios correspondientes en las otras versiones de la fórmula. Nosotros usamos el símbolo "±" con E por ser, quizá, una versión más accesible de las fórmulas.

3 Técnicamente, se distingue entre diversos tipos de formas alternas, por ejemplo, formas estrictamente paralelas, equivalentes τ y esencialmente equivalentes τ. Las diferencias entre estas formas tienen implicaciones para algunos temas de psicometría avanzada. Para ahondar en estas distinciones, se puede consultar Lord y Novick (1968), Feldt y Brennan (1989) y Nunnally y Bernstein (1994).

4 Thorndike (1982) sugirió un procedimiento para evitar esta suposición, pero su sugerencia no se usó mucho.

CAPÍTULO 5
Validez

Objetivos

1. Comparar las definiciones "estándar" y "refinada" de la validez de las pruebas.

2. Usar los conceptos de subrepresentación del constructo y varianza irrelevante para el constructo para definir la validez de las pruebas.

3. Identificar las tres categorías tradicionales para describir la evidencia de la validez.

4. Definir validez aparente.

5. Definir validez de contenido y discutir sus usos típicos.

6. Definir validez de criterio y discutir sus tres usos típicos.

7. Discutir los efectos de la confiabilidad de la prueba y del criterio sobre la validez de criterio.

8. Ilustrar el uso de la correlación múltiple para demostrar la validez incremental.

9. Definir validez convergente y discriminante y usar la matriz multirrasgo-multimétodo.

10. En el contexto de la validez de criterio, ilustrar los conceptos de positivos falsos, negativos falsos, índice base, selectividad y especificidad.

11. Definir validez de constructo y dar varios ejemplos donde se aplique este concepto.

12. Describir el objetivo del análisis factorial.

13. Definir el papel de estudiar los procesos de respuesta en la validez de constructo.

14. Discutir el significado de la validez consecuencial.

15. Discutir el significado de generalización de la validez y metaanálisis al considerar la validez de una prueba.

Introducción

Empecemos con estas situaciones y preguntas prácticas relacionadas con la validez de las pruebas.

- El Colegio Ivy emplea el *Western Admissions Test* (WAT [Prueba de Admisión del Oeste]) para elegir a los aspirantes que pueden tener éxito en sus estudios. ¿Qué tipo de evidencia debería buscarse para determinar si el WAT cumple su propósito?
- El Dr. Arias contempla usar el *Scranton Depression Inventory* [Inventario Scranton de Depresión] para ayudar a identificar la gravedad de la depresión y, en especial, para distinguir entre depresión y ansiedad. ¿Qué evidencia debería emplear el Dr. Ally para decidir si la prueba hace lo que él espera?
- El recién publicado *Disgnostic Wonder Test* [Prueba Diagnóstica Maravillosa] promete identificar niños con problemas de aprendizaje de las matemáticas. ¿Cómo sabremos si la prueba hace esto o es sólo una prueba de capacidad general publicitada con mucha habilidad?
- Miguel revisa un informe narrativo de sus puntuaciones en el *Nifty Personality Questionnaire* (NPQ [Cuestionario Sensacional de Personalidad]). El informe dice que él es excepcionalmente introvertido y poco curioso frente al mundo que lo rodea. ¿Miguel puede confiar en estas afirmaciones? ¿O debe hacer caso omiso de ellas como si se tratara de lectura de la mano en un lugar mágico?
- Un sistema escolar quiere usar una batería de aprovechamiento que medirá el grado en que los estudiantes aprenden lo que marca el programa de la escuela. ¿Cómo debe proceder el sistema escolar para revisar las pruebas de aprovechamiento disponibles?

Refinando la definición de validez

Todas estas preguntas se relacionan con la validez de las pruebas. En este capítulo, refinaremos nuestra forma de pensar sobre este tema y examinaremos métodos empleados para responder a estas preguntas. La definición habitual de **validez** es el grado en que la prueba mide lo que pretende medir. Citamos esta definición en el capítulo 1 al discutir las preguntas fundamentales en el campo de las pruebas psicológicas. Esta definición se usa, a menudo, en los libros introductorios de psicología. En ese nivel elemental, propusimos plantear esta pregunta: ¿esta prueba es válida? Sin embargo, ahora que tenemos la oportunidad de tratar el tema de la validez con mayor detalle, necesitamos refinar nuestras ideas reformulando la pregunta de tres maneras.

> Lo que se evalúa en la validez de una prueba, es la interpretación de las puntuaciones de la prueba, requerida para los propósitos que se le pretende dar, no la prueba en sí misma. Cuando las puntuaciones se usan o se

interpretan en más de un modo, cada interpretación tiene que ser validada.

Standards... (AERA, APA, & NCME, 2013)

Primero, debemos señalar que es impreciso referirse a la validez de una prueba; lo que necesitamos es determinar la validez de la puntuación de una prueba cuando se usa con un propósito específico. Con mayor exactitud aún, debemos referirnos a la interpretación de una puntuación con un propósito o uso específico. Notemos que en los escenarios esbozados al principio de este capítulo siempre declaramos un propósito para la prueba. El uso de una puntuación puede ser apropiado para cierto propósito, pero no para otro, por lo que no podemos definir la validez de la puntuación de una prueba en lo abstracto, sino sólo con respecto a un uso específico. Así, no debemos hacernos preguntas como: ¿el Rorschach es válido? O ¿el SAT es válido? En cambio, las preguntas deben plantearse así: ¿el índice de Depresión del Rorschach es válido para identificar la gravedad de una depresión? O ¿la puntuación de matemáticas del SAT es válida para predecir el GPA de un estudiante al final del primer año en la universidad?

Segundo, la validez es una **cuestión de grado**, no de todo o nada. Algunas pruebas pueden no tener validez para propósitos específicos; de hecho, es probable que no existan puntuaciones de alguna prueba que sean perfectamente válidas para cierto propósito. La mayoría de las puntuaciones que usamos tienen cierto grado de validez, que puede ser leve, moderada o considerable. Nuestro interés determinará el grado de validez; desde el punto de vista práctico, queremos saber si la validez es suficiente para hacer buen uso de la prueba. Así, refinaremos más nuestra pregunta del siguiente modo: ¿en qué grado el Índice de Depresión de Rorschach es válido para determinar la gravedad de una depresión? O ¿en qué grado la puntuación de matemáticas del SAT es válida para predecir el GPA de un alumno de primer año?

Tercero, debemos distinguir entre validez y exactitud de las normas de una prueba. Es muy posible tener una prueba con una buena validez, pero también con normas bastante inexactas. Cuando esto ocurre, algunas personas concluyen, de manera errónea, que la prueba no es válida. Consideremos los siguientes escenarios. Las pruebas A y B son de "CI" y se usan para predecir el GPA en la universidad; ambas pruebas tienen una correlación de .65 con el GPA. En los dos casos, el promedio del GPA es 3.0; en la prueba A, el CI promedio es 110, mientras que en la prueba B, el promedio es 80. El usuario podría concluir que la prueba B "no es válida", porque no tiene sentido pensar que estudiantes con un CI promedio de 80 puedan tener un GPA de 3.0 en promedio. Sin embargo, como mencionamos, el coeficiente de validez (r = .65) es el mismo en ambas pruebas, por lo que tienen la misma validez para predecir el GPA. El problema está en la exactitud de las normas, no en la validez de la prueba. Desde luego, lo contrario también puede ser cierto, es decir, que una prueba pueda tener normas excelentes, pero poca o ninguna validez.

Subrepresentación del constructo y varianza irrelevante para el constructo

Conforme formalizamos nuestro tratamiento de la validez, dos términos técnicos serán de ayuda para nuestras ideas. Pero antes de introducirlos, consideremos la superposición entre el *constructo* que deseamos medir y la *prueba* que esperamos que lo mida. El **constructo** es un rasgo o característica; por ejemplo, el constructo podría ser depresión o capacidad de razonamiento matemático. Podemos tener un cuestionario sencillo de 20 reactivos para medir depresión y una prueba de 50 reactivos de opción múltiple para medir el razonamiento matemático. Representamos la relación entre constructo y prueba superponiendo formas geométricas como en la figura 5-1; la superposición entre constructo y prueba representa la validez: medir lo que queremos medir. La parte del constructo que *no* está cubierta por la prueba es lo que llamamos **subrepresentación del constructo**. El constructo de interés no está cubierto en su totalidad por la prueba. Por otro lado, la prueba, además de medir una parte del constructo de interés, puede medir algunas características diferentes de las que queremos medir; esta "otra" medición se llama **varianza irrelevante para el constructo**.

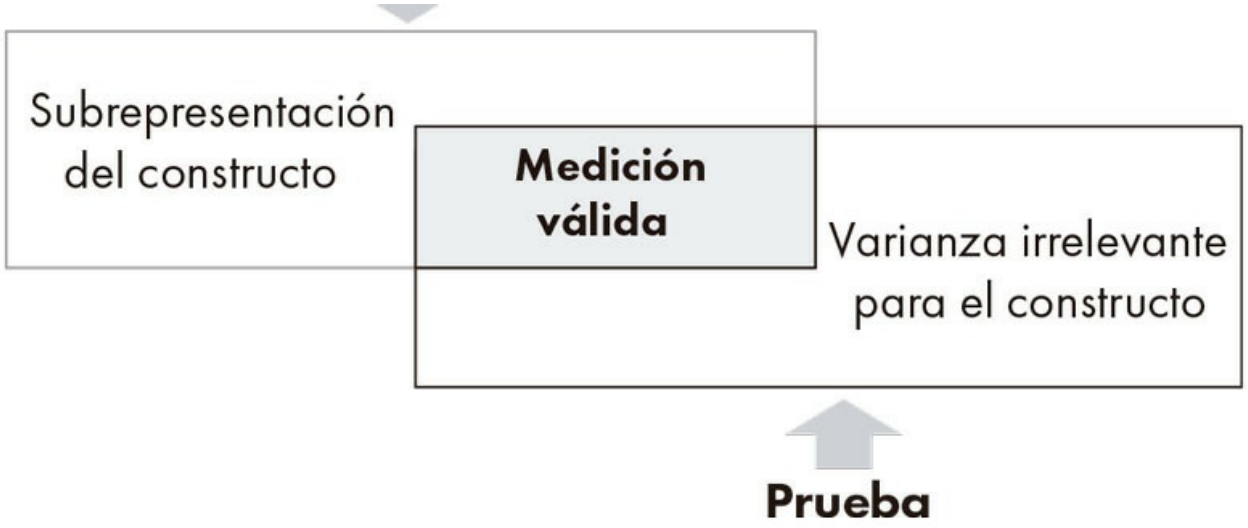


Figura 5-1. Representación geométrica de los conceptos de subrepresentación del constructo y varianza irrelevante para el constructo.

Primero, consideremos algunos ejemplos concretos para, luego, examinar cómo pueden representarse de manera gráfica. Supongamos que el concepto de depresión consta de tres componentes: cognitivo (pensamientos acerca de la depresión), emocional (sentirse deprimido) y conductual (hacer o no hacer cosas sintomáticas de la depresión). Nuestro cuestionario puede hacer un trabajo excelente abordando los componentes cognitivo y emocional, al mismo tiempo que deja fuera la información sobre el componente conductual. Así, el constructo completo de depresión es subrepresentado por la prueba, en particular, por la omisión de su componente conductual. Este análisis supone que los tres componentes son, al menos parcialmente, independientes y no sólo nombres diferentes de la misma característica. También puede ocurrir que, en cierto grado, las

puntuaciones del cuestionario reflejen una tendencia en las respuestas hacia la deseabilidad social. Esto no es lo que queremos medir, por lo que este aspecto de las puntuaciones es varianza irrelevante para el constructo.

Apliquemos estos conceptos a la prueba de razonamiento matemático; esperamos que este constructo se manifieste en la capacidad para resolver problemas convencionales y novedosos. Sin embargo, los reactivos de la prueba sólo incluyen problemas convencionales, por lo que la parte del constructo relacionada con los problemas novedosos está subrepresentada. (Suponemos que el razonamiento en problemas novedosos no tiene una correlación perfecta con el razonamiento en problemas convencionales. Si su correlación fuera perfecta, o casi perfecta, no haría ninguna diferencia el tipo de problemas que usamos.) Además, la prueba requiere un nivel muy alto de capacidad de lectura, pero no queremos que la prueba sea de lectura. La parte de las puntuaciones determinada más por esta capacidad que por la de razonamiento matemático constituye la varianza irrelevante para el constructo.

Podemos tener una variedad infinita de relaciones entre constructo y prueba. Ésta puede cubrir gran parte del constructo, y además tener mucha varianza irrelevante; o la prueba puede tener poca varianza irrelevante, pero cubrir una parte mínima del constructo. La figura 5-2 muestra varias de estas posibilidades. Desde luego, lo ideal sería la superposición total del constructo y la prueba, pero lo más común en la práctica es no alcanzar este ideal. Las nociones de subrepresentación del constructo y varianza irrelevante para el constructo serán muy útiles cuando examinemos diferentes métodos de investigar la validez de las pruebas.

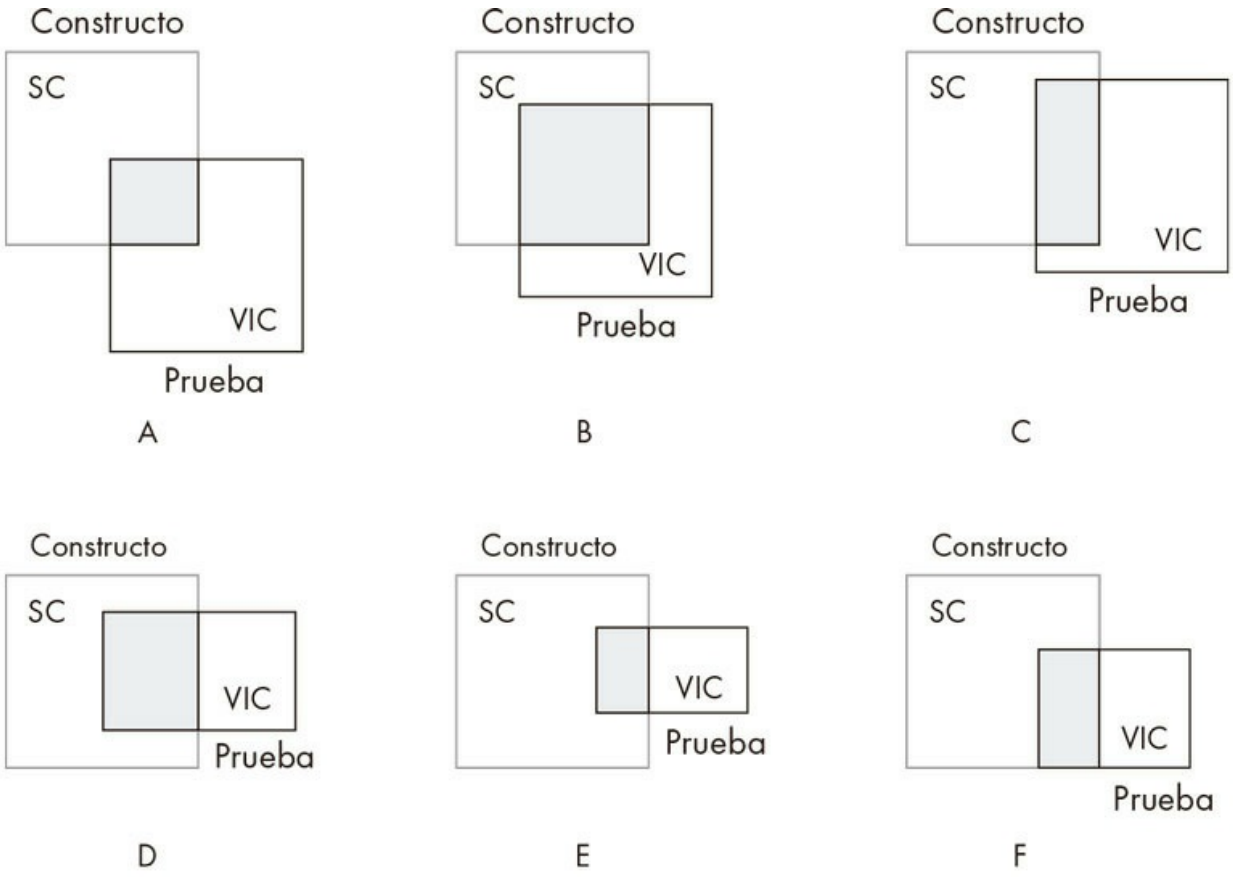


Figura 5-2. Ilustración de los grados variables de la subrepresentación del constructo y la varianza irrelevante para el constructo.

> La subrepresentación del constructo se refiere al grado en que una prueba no logra capturar aspectos importantes del constructo... La irrelevancia para el constructo se refiere al grado en que las puntuaciones de la prueba son afectadas por procesos externos al constructo para el que fue pensada la prueba.
>
> *Standards...* (AERA, APA, & NCME, 2013)

¡Inténtalo!

¿Cuál de los ejemplos de la figura 5-2 muestra el grado más alto de validez? ¿Qué ejemplo muestra el grado más alto de varianza irrelevante para el constructo?

La cuestión básica

La cuestión básica es proporcionar evidencia de que las puntuaciones de una prueba son indicios del rasgo o constructo de interés. Esta evidencia constituye el fundamento de la interpretación de las puntuaciones de la prueba. Nuestro tratamiento de la validez de las pruebas toma en cuenta los tipos de evidencia que parecen ser convincentes en relación con esta cuestión básica. Revisamos los tipos de evidencia que se requiere presentar para

establecer la validez de una prueba y, por fuerza, también discutimos el cuidado especial que se debe tener al interpretar dicha evidencia e introducir los términos especializados que los psicólogos han desarrollado para este tema. Existe un consenso general de que la validez es la característica más importante de una prueba. Buenas normas, confiabilidad alta y otras características deseables son importantes, pero no tienen sentido si no hay una buena validez.

> La validez es... lo más importante que hay que considerar al elaborar una prueba y evaluarla.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Clasificaciones tradicionales y nuevas de los tipos de evidencia de la validez

Hay un sistema tradicional tripartito para clasificar los tipos de evidencia de la validez que está muy arraigado en la literatura psicométrica. *Standards* de 1999 abandonó, en parte, este sistema en favor de una representación más diversificada de los tipos de evidencia; en su edición más reciente, continuó con este nuevo sistema prácticamente sin cambios. Sin embargo, el sistema tradicional ha demostrado tener una vitalidad sorprendente. Aquí presentamos tanto el sistema tradicional como el más reciente; en el cuadro 5-1 se bosquejan. El lector contemporáneo debe estar familiarizado con la terminología de ambos sistemas.

Cuadro 5-1. Esbozo del sistema tradicional y el más reciente para clasificar los tipos de validez

Clasificación tradicional	**Clasificación reciente**
De contenido	De contenido
De criterio	Relaciones con otras variables
Concurrente	Convergente y discriminante
Predictiva	Relaciones con pruebas-criterios
De constructo	Procesos de respuesta
	Estructura interna
	Consecuencias

Trataremos cada una de estas categorías en secciones subsiguientes de este capítulo, pero primero presentamos, en el cuadro 5-1, una comparación de la terminología que se usa en el sistema tradicional y en el más reciente. La validez de contenido tiene, en gran parte, el mismo significado en los dos sistemas. En el sistema tradicional, la validez de criterio corresponde de manera muy cercana a la nueva categoría "relaciones con otras variables", sobre todo a la subcategoría "relaciones con otras pruebas-criterio". Los conceptos de validez convergente y discriminante se reflejan en el sistema tradicional, pero son mucho más explícitos en el sistema más reciente. Asimismo, los estudios de los procesos de respuesta y estructura interna se representan en el sistema tradicional, pero bajo la categoría general de validez de constructo, la cual no es una categoría principal en el sistema más reciente aunque las ediciones recientes de *Standards* están permeadas por

la noción de validez de constructo.

Como señalamos más adelante en este capítulo, las "consecuencias" son todo un nuevo tema.

En las siguientes secciones de este capítulo, cubriremos todos los elementos importantes de ambos sistemas; sin embargo, aquí hacemos hincapié, al igual que al final del capítulo, en que establecer la validez no es cuestión de pasar lista a los elementos del esquema de clasificación, sino que implica presentar un arreglo integrado multifacético de evidencia con respecto a la interpretación adecuada de la puntuación de una prueba. Además de *Standards*, una fuente esencial sobre todos los intentos de validación en el contexto de la contratación de empleados es *Principles for the Validation and Use of Personnel Selection Procedures* [Principios de validación y uso de procedimientos de selección de personal], preparado por la *Society for Industrial and Organizational Psychology, Inc.* [Sociedad de Psicología Industrial y Organizacional] (SIOP, 2003).

La cuestión de la validez aparente

Cuando los psicólogos se refieren a la validez de las pruebas, hablan de una demostración empírica de que una prueba mide lo que se propone medir y, de manera más específica, de que las puntuaciones de la prueba pueden interpretarse de manera significativa con algún objetivo particular. Contrastamos este enfoque empírico con la **validez aparente**, la cual se refiere a si la prueba *tiene la apariencia de* medir el constructo meta. La validez aparente tiene defensores y detractores; estos últimos se burlan de ella, porque a menudo se usa como sustituto de la demostración empírica de la validez. Puede ser seductora y engañosa. El autor de una prueba puede decir: "La inspección de los reactivos de *Scranton Anxiety Test* [Prueba Scranton de Ansiedad] indica con claridad que la prueba mide las principales facetas de la ansiedad". Ante la ausencia de cualquier otro apoyo, esta afirmación no ayuda. Por otro lado, los defensores de la validez aparente hacen notar que trabajamos con personas reales en el mundo real. Debemos tener, sin duda, una validez demostrada empíricamente; sin embargo, en la mayoría de las circunstancias, también es útil si la prueba tiene la apariencia de una medida válida.

Hacemos la siguiente recomendación acerca de la validez aparente. Ésta nunca puede ser un sustituto de la validez empírica, es decir, no podemos nada más ver una prueba y saber si tiene algún grado de validez. Sin embargo, la validez aparente puede ser útil; si dos pruebas tienen una validez empírica equivalente, suele ser preferible usar la que tiene mejor validez aparente. Cuando se construye una prueba, es prudente buscar la validez aparente, pero nunca a expensas de la validez empírica. Quizá más importante, necesitamos tener siempre presente la diferencia entre validez aparente y validez demostrada empíricamente.

Validez de contenido

La **validez de contenido** se ocupa de la relación entre el contenido de una prueba y algún dominio bien definido de conocimiento o conducta. Para que una prueba tenga validez de contenido, debe haber una buena correspondencia entre el contenido de la prueba y el contenido del dominio pertinente. La obtención de la validez de contenido a menudo implica la noción de muestreo, es decir, el contenido de la prueba cubre una muestra representativa de todos los posibles contenidos del dominio. Esto no siempre debe ser así, pues la prueba puede cubrir todo el material del dominio; sin embargo, lo más habitual es que el dominio sea demasiado grande para que se pueda cubrir todo. Es entonces cuando nos apoyamos en el muestreo. La validez de contenido tiene dos aplicaciones primarias: pruebas educativas de aprovechamiento y pruebas de reclutamiento laboral. En cada una de estas áreas, hay un cuerpo bien definido de contenido. Queremos determinar el grado en que el contenido de la prueba se ajusta al contenido del área educativa o puesto de trabajo pertinentes.

Aplicación en las pruebas de aprovechamiento

Por lo general, la validez de contenido se considera el tipo más importante de validez para las pruebas de aprovechamiento. El propósito habitual de estas pruebas es determinar el grado de conocimiento sobre algún material. El cuadro 5-2 presenta ejemplos de los materiales que pueden ser el objetivo de una prueba de aprovechamiento.

El proceso de establecer la validez de contenido empieza con una definición cuidadosa del contenido que se desea cubrir. Este proceso suele resultar en un cuadro de especificaciones o un anteproyecto. Consideremos algunas de las entradas del cuadro 5-2; el cuadro de especificaciones para "química de nivel bachillerato" puede surgir de examinar el contenido de los cinco libros más usados en este campo. El cuadro de especificaciones para "capítulo 5 de este libro" puede surgir de la lista de objetivos y de palabras clave que aparecen al inicio y al final del capítulo, respectivamente. "Conceptos matemáticos de 1 a 3 grado" pueden definirse con las guías curriculares de distintos estados. Lo más frecuente es que los documentos escritos sirvan de base para el cuadro de especificaciones. El cuadro 5-3 cita afirmaciones sobre las bases del contenido de dos pruebas estandarizadas de aprovechamiento: *Major Field Tests*, pruebas de aprovechamiento de nivel universitario en 14 disciplinas, y *Stanford Achievement Test*, batería multinivel para los grados K-12. Podemos notar cómo las referencias al contenido definen la orientación de las pruebas.

Cuadro 5-2. Ejemplos de campos de conocimiento como objetivo de las pruebas de aprovechamiento

Conceptos matemáticos de 1 a 3 grado

Química de nivel bachillerato
Primer curso de pruebas psicológicas
Capítulo 5 de este libro
Lecciones de geografía de la clase de la maestra Vásquez de la semana pasada
Ortografía típica de escuelas primarias
Adición, sustracción, multiplicación y división en situaciones numéricas
Historia de la Guerra Civil
Habilidades básicas de escritura

Cuadro 5-3. Afirmaciones del propósito de la prueba orientadas hacia la validez de contenido

"... Major Field Tests [Pruebas de los Campos Principales] son evaluaciones de resultados... completas para nivel de licenciatura diseñadas para medir el conocimiento y la comprensión críticos de los estudiantes en un campo importante de estudio. Los Major Field Tests van más allá de la medición de conocimiento objetivo, pues ayudan a evaluar la capacidad del estudiante para analizar y resolver problemas, comprender relaciones e interpretar material de su campo de estudio." (Educational Testing Service, 2012)

"La serie de Stanford Achievement Test [Prueba Stanford de Aprovechamiento]… evalúa el aprovechamiento escolar del alumno en lectura, matemáticas, ortografía, lenguaje, ciencia, ciencias sociales y comprensión auditiva... Los reactivos incluidos en Stanford 10 reflejan la extensa revisión de los estándares de enseñanza nacionales y estatales, los currícula de contenido específico y las tendencias educativas tal como fueron desarrolladas por organizaciones educativas profesionales a nivel nacional." (Harcourt Educational Measurement, 2003, p. 5)

En muchos casos, un área de contenido se representa por medio de *un cuadro de especificaciones de dos vías*. La primera dimensión del cuadro cubre los temas del contenido, mientras que la segunda representa los procesos mentales, como el conocimiento objetivo, comprensión de conceptos y capacidad para aplicar o sintetizar material.

El esquema más conocido para representar los procesos se denomina taxonomía de Bloom. Ésta es una ramificación del trabajo de Benjamin Bloom y sus colegas, quienes elaboraron tres taxonomías o esquemas de clasificación: una en el dominio cognitivo (Bloom, 1956), otra en el dominio afectivo (Krathwohl, Bloom, & Masia, 1964) y una más, poco usada, en el dominio psicomotor (Harrow, 1972). El cuadro 5-4 bosqueja las principales categorías de la taxonomía cognitiva, la cual es la más citada de las tres y, también, la más pertinente para nuestra discusión sobre la validez de contenido de las pruebas de aprovechamiento. Aunque a veces se usa la taxonomía cognitiva completa, se suele reducir a tres categorías principales, de las seis que la integran, con el mismo nombre: taxonomía de Bloom. Los esfuerzos por validar las distinciones de la taxonomía cognitiva de Bloom, es decir, para mostrar que las distintas categorías representan procesos mentales relativamente distintos, han fracasado (Kreitzer & Madaus, 1994; Seddon, 1978). No obstante, esta taxonomía o una variación de ella se encuentran con frecuencia en las discusiones sobre la validez de contenido.

Cuadro 5-4. Principales categorías de la taxonomía de Bloom para el dominio cognitivo

Conocimiento	Comprensión	Aplicación
Análisis	Síntesis	Evaluación

Nosotros usamos un sistema reducido tipo Bloom en el cuadro 5-5 para ilustrar un cuadro de dos vías con las especificaciones sobre el contenido del capítulo 4 de este libro: confiabilidad. Las entradas en las casillas del cuadro 5-5 muestran el peso relativo asignado a cada casilla en forma de porcentaje.

Cuadro 5-5. Ejemplo de un cuadro de dos vías con las especificaciones del contenido basadas en el material del capítulo 4 de este libro: confiabilidad

Contenido	**Proceso**			
	Hechos	**Conceptos**	**Aplicaciones**	**Total**
Fuentes que atentan contra la confiabilidad	5	5	—	10
Método de test-retest	3	5	5	13
Confiabilidad interjueces	3	3	3	9
Consistencia interna	5	10	5	20
Error estándar	5	5	5	15
Pruebas con referencia a un criterio	3	3	2	8
Teoría de la generalizabilidad	2	3	—	5
Factores que afectan *r*	5	5	10	20
Total	**31**	**39**	**30**	**100**

Por ejemplo, cerca de 10% del contenido se ocupa de conceptos relacionados con la consistencia interna; por lo tanto, cerca de 10% de los reactivos debe abordar conceptos relacionados con la consistencia interna; en una prueba de 50 reactivos, significaría incluir cinco reactivos sobre este tema. Si hubiera sólo un reactivo sobre este tema, o 20, la prueba tendría una validez de contenido pobre. En términos de los totales marginales del cuadro 5-5, esperaríamos que cerca de 20% de reactivos (10 en una prueba de 50 reactivos) abordara el tema "factores que afectan *r*".

Después de preparar un cuadro de especificaciones sobre un área de contenido, determinamos la validez de contenido de una prueba contrastando su contenido con el cuadro de especificaciones. Esto suele hacerse reactivo por reactivo; este análisis debe mostrar a) áreas de contenido que la prueba no cubre y b) reactivos que no se ajustan a las especificaciones del contenido. Podemos notar que estas dos áreas corresponden en gran medida a las nociones de subrepresentación del constructo y varianza irrelevante para el constructo de las que hablamos antes.

La última descripción se aplica al determinar la validez de contenido de una prueba existente; se usa un proceso similar cuando una prueba de aprovechamiento se está elaborando. Sin embargo, ahora nosotros preparamos los reactivos de la prueba de

manera específica para que se ajusten al anteproyecto del contenido. En las páginas 135-137a» del capítulo 6 se describe el proceso de elaboración de pruebas con mayor detalle.

> Los creadores de pruebas a menudo trabajan a partir de una especificación del dominio de contenido. En estas especificaciones se describe de manera cuidadosa y detallada el contenido, a menudo con una clasificación de las áreas de contenido y los tipos de reactivos.
>
> *Standards...* (AERA, APA, & NCME, 2013)

Dada la manera en que determinamos la validez de contenido de una prueba de aprovechamiento, podría pensarse que podemos resumir los resultados de manera numérica, es decir, que podríamos expresar el porcentaje del dominio cubierto por los reactivos y el porcentaje de reactivos que no reflejan el dominio. En la práctica, esto se hace rara vez; en su lugar, después de ajustar el contenido de la prueba al dominio, se emite un juicio acerca de la validez de contenido: suficiente o insuficiente.

Validez instruccional

Una aplicación especial de la validez de contenido es la noción de validez instruccional, también conocida como validez curricular. Mientras que la validez de contenido pregunta si el contenido de la prueba se ajusta bien a cierto contenido, la **validez instruccional** pregunta si el contenido ha sido, en verdad, enseñado. Para que una prueba tenga validez instruccional, debe haber evidencia de que el contenido se cubrió de manera adecuada en un programa de enseñanza. A veces llamamos a esto "oportunidad de aprender". En algunos contextos, preguntamos si los estudiantes que responden la prueba en realidad han sido expuestos al material que cubre la prueba.

El concepto de validez instruccional se aplica primordialmente a las pruebas de aprovechamiento educativo. Consideremos el tema de la raíz cuadrada; éste puede aparecer en la guía curricular de la escuela y en el libro de matemáticas que se usa en la escuela. Por lo tanto, la prueba de aprovechamiento de la escuela incluye reactivos sobre la raíz cuadrada. Ésa es una buena validez de contenido. Supongamos, sin embargo, que ninguno de los maestros de la escuela cubrió ese tema en clase ni en las tareas para casa. Entonces los reactivos sobre la raíz cuadrada no tienen validez instruccional: no hubo "oportunidad de aprender" acerca de la raíz cuadrada.

La noción de validez instruccional no está bien establecida como algo distinto de la validez de contenido. Standards no incluye el término validez instruccional, pero hay una pequeña discusión sobre el concepto de oportunidad de aprender. En efecto, la noción de validez instruccional hace referencia simplemente al "contenido bien definido" que es en verdad enseñado más que el que se supone que debe haberse enseñado. Ésta es una distinción útil, pero no introduce una validez por completo nueva; no obstante, el término **validez instruccional o validez curricular** ha aparecido. Fue un concepto destacado en un famoso caso en la corte, Debra P vs. Turlington, que veremos en el capítulo 16.

Aplicación en las pruebas de reclutamiento

La segunda aplicación de la validez de contenido es en las pruebas de reclutamiento o selección de personal. Las nociones esenciales son las mismas que las de las pruebas de aprovechamiento educativo. En las pruebas de reclutamiento, el dominio de contenido consiste en los conocimientos y habilidades requeridos para un trabajo específico. Cuando se construye la lista de especificaciones sobre el trabajo, es habitual restringir la lista a los conocimientos y habilidades que se requieren, específicamente para el nivel inicial. Los factores como motivación y características de personalidad no suelen incluirse, pues pueden evaluarse en el proceso de selección mediante otras pruebas que aquí no discutiremos. Además, estas otras pruebas tendrían que validarse siguiendo métodos diferentes de la validez de contenido. El proceso de desarrollar una lista de conocimientos y habilidades necesarios para un trabajo a menudo se denomina **análisis de puesto**. Después de hacer el análisis de puesto, ajustamos el contenido de la prueba al contenido del puesto. Al igual que con las pruebas de aprovechamiento, podemos ajustar una prueba existente a un conjunto de especificaciones de un puesto, o podemos construir una nueva prueba que se ajuste a dichas especificaciones.

Aunque hay muchas similitudes al aplicar la validez de contenido a las pruebas de aprovechamiento y reclutamiento, existen dos diferencias interesantes. Primero, en el caso de las pruebas de aprovechamiento, los documentos impresos, como libros de texto o guías curriculares, por lo general sirven como base de las especificaciones del contenido, mientras que en las de reclutamiento, a menudo un panel de expertos desarrolla las especificaciones. Se puede encontrar una descripción detallada de este proceso en Knapp y Knapp (1995); estos autores también presentan una útil revisión de casos de tribunal relacionados con la necesidad de validez de contenido y un análisis de puesto adecuado. Segundo, aunque rara vez se usa cifra del porcentaje de acuerdo en las pruebas de aprovechamiento, en las de reclutamiento sí se usa para su evaluación. Lawshe (1978) presentó una metodología para expresar el porcentaje del contenido de la prueba que un panel de expertos juzgó esencial para el desempeño en un trabajo; su resultado lo denominó razón de validez de contenido. Schmidt, Ones y Hunter (1992) y Borman, Hanson y Hedge (1997) presentaron una revisión útil de la investigación relacionada con el análisis de puesto. Raymond (2001, 2002) aplicó este concepto a los exámenes de certificación y licencias.

> La evidencia basada en el contenido también puede venir de juicios de expertos acerca de la relación entre partes de la prueba y el constructo. Por ejemplo, al desarrollar una prueba para conceder una cédula profesional, pueden especificarse las principales facetas que son pertinentes para el propósito de la profesión que se regula, y se les puede pedir a los expertos de esa profesión que clasifiquen los reactivos de la prueba de acuerdo con las categorías definidas por dichas facetas.
>
> *Standards...* (AERA, APA, & NCME, 2013)

Validez de contenido en otras áreas

Como señalamos antes, la validez de contenido tiene su principal aplicación en las pruebas de aprovechamiento educativo y de reclutamiento. Su aplicación en otras áreas, por ejemplo, la inteligencia y la personalidad, es limitada, porque pocas áreas son susceptibles de hacer especificaciones claras de los dominios que se deben cubrir. Por ejemplo, ¿cuál es el contenido de la inteligencia o la extroversión? Aunque podemos tener definiciones sencillas de estos constructos, es difícil especificar un bosquejo detallado de lo que comprenden. De ahí que la validez de contenido no se aplique con claridad a ellos. Sin embargo, en algunos casos, la validez de contenido puede tener un uso limitado en estas áreas; por ejemplo, puede ser útil para mostrar que una prueba diseñada para medir cierto trastorno de personalidad cubre todos los rasgos especificados de dicho trastorno en el *DSM Manual diagnóstico y estadístico de los trastornos*. Tratamos justo este punto en algunas de las pruebas que presentamos en el capítulo 13: Instrumentos y métodos clínicos. Sin embargo, por lo general, nos apoyamos en otros métodos para demostrar la validez de dichas pruebas.

Problemas con la validez de contenido

Establecer la validez de contenido siempre parece un proceso muy sencillo. En términos conceptuales, es muy básico: especificar el contenido del dominio y, luego, revisar qué tan bien se ajusta la prueba a este contenido. Sin embargo, en la práctica, el proceso casi siempre resulta ser mucho más complicado, lo cual se deriva de tres fuentes. Primero, excepto en algunos casos muy sencillos, a menudo es difícil obtener una especificación clara del dominio de contenido. Consideremos los ejemplos del cuadro 5-2. Dijimos que el contenido de "conceptos matemáticos en los grados 1 a 3" podía determinarse revisando las guías curriculares de varios estados, pero estas guías difieren un poco de un estado a otro. Supongamos que revisamos las guías de cinco estados; tres pueden incluir conocimiento de las unidades métricas en los grados 1 a 3, pero otras dos pueden posponer este tema hasta el grado 4. ¿Cómo manejamos esto? Al especificar el contenido del "capítulo 5 de este libro", qué nivel de profundidad del conocimiento queremos: ¿un conocimiento pasajero de los temas principales o una comprensión completa de cada detalle? Podemos hacer preguntas similares acerca de los conocimientos y habilidades enumeradas en las especificaciones de una prueba de reclutamiento.

La segunda dificultad para aplicar la validez de contenido proviene de juzgar qué tan bien los reactivos de la prueba cubren los elementos de las especificaciones del contenido. Los reactivos con una clasificación común pueden variar mucho en las habilidades que demandan. Consideremos los ejemplos del cuadro 5-6; muchos reactivos diferentes se aplican a una categoría de contenido como "operaciones de multiplicación básica". ¿Todos estos reactivos son igual de apropiados? ¿Todos miden la categoría de contenido igual de bien? Probablemente no. El ejemplo de una categoría de contenido usada aquí –operaciones de multiplicación básica– es sencillo. Imaginemos cuánto más complicada se vuelve la situación con un tema más complejo, como el conocimiento de

la Guerra Civil o las habilidades básicas de escritura. En una lista de contenido de la prueba, todos los reactivos del cuadro 5-6 podrían categorizarse como "operaciones de multiplicación básica". La persona que juzga la validez de contenido debe examinar los reactivos reales de la prueba y no basarse sólo en una lista de categorías. En el análisis final, la validez de contenido requiere un juicio y no sólo pasar lista a los elementos que incluye.

Cuadro 5-6. Diversos reactivos que corresponden a una sola categoría de contenido

Contenido meta: Operaciones de multiplicación básica							
Posibles reactivos de la prueba							
1. 5 × 4 = _____							
2. 5 × [ ] = 20							
3. 5 × 4 =				(a) 9	(b) 20	(c) 25	(d) 7
4. 5 × [ ] = 20	[ ] =			(a) 15	(b) 4	(c) 5	(d) 25
5. Jack compró cuatro dulces a 5 pesos cada uno. ¿Cuánto dinero gastó?							
6. Jack compró cuatro dulces a 5 pesos cada uno. ¿Cuánto dinero gastó?							
	(a) 9	(b) 20	(c) 25	(d) Ninguna de las anteriores			
7. Jack pagó 20 pesos por cuatro dulces. ¿Cuánto le costó cada uno?							

Una tercera dificultad con la validez de contenido es que no hace referencia en ningún sentido al desempeño real en la prueba. Todos los demás métodos para determinar la validez se refieren, al menos en cierto sentido, al desempeño empírico. Así, la validez de contenido nos deja desanclados del mundo real de la interacción entre el examinado y la prueba.

¡Inténtalo!

Escribe dos reactivos más que se apliquen al contenido meta enumerado en el cuadro 5-6: "operaciones de multiplicación básica".

Validez referida al criterio

La característica esencial de la validez referida al criterio es establecer la *relación entre el desempeño en la prueba y algún otro criterio* que se considera un indicador importante del constructo de interés. Hay tres aplicaciones comunes de la validez de criterio en dos contextos generales. En todos los casos, tratamos de establecer la relación entre el desempeño en la prueba y su condición según algún otro criterio.

Los dos contextos generales de la validez de criterio son la validez predictiva y la validez concurrente. En la **validez predictiva**, la prueba busca predecir el estatus en algún criterio que será alcanzado en el futuro; por ejemplo, podemos usar una prueba de ingreso a la universidad, aplicada en el último año de bachillerato, para predecir el GPA al final del primer año en la universidad. O podemos usar un inventario de personalidad para predecir la probabilidad de un intento de suicido en algún momento futuro. En la **validez concurrente**, verificamos la concordancia entre el desempeño en la prueba y el estatus actual en alguna otra variable; por ejemplo, podemos determinar la relación entre el desempeño en una prueba estandarizada de aprovechamiento y una prueba hecha por el profesor, para lo cual ambas deben aplicarse casi al mismo tiempo. O podemos determinar la relación entre la puntuación de una prueba de depresión y la valoración del clínico acerca del nivel actual de depresión. La diferencia entre validez predictiva y concurrente es estrictamente temporal en relación con la variable criterio. Por lo demás, los dos conceptos son lo mismo.

> A lo largo de la historia, dos diseños, a menudo llamados predictivo y concurrente, se han distinguido por evaluar las relaciones entre prueba y criterio.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Las tres aplicaciones comunes de la validez de criterio implican el uso de a) un criterio externo y factible que defina el constructo de interés, b) contrastes grupales y c) otra prueba. En lo fundamental, estos tres enfoques se reducen a lo mismo; sin embargo, tienen algunas diferencias prácticas, así que los trataremos por separado.

Criterio externo y factible

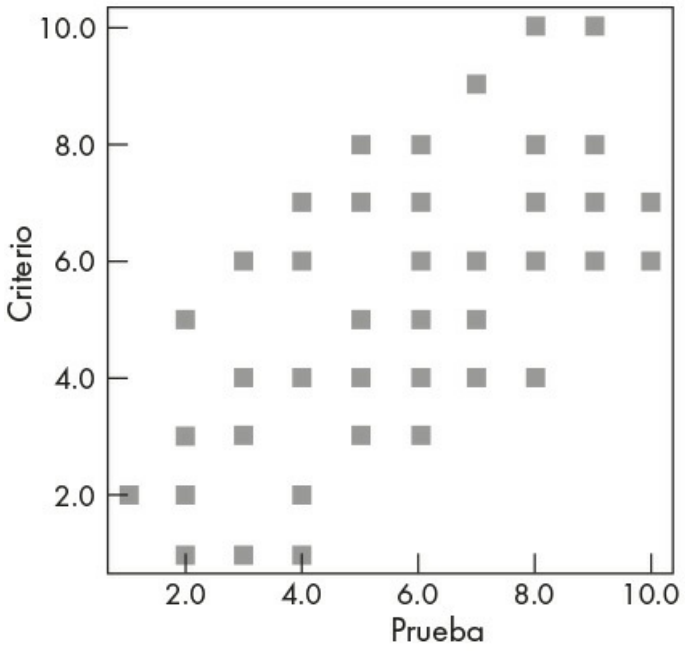
En algunas circunstancias, tenemos un **criterio externo** que proporciona una definición factible del constructo de interés. El criterio externo es aquello sobre lo que nos gustaría tener información; la pregunta natural es: si en verdad queremos información de un criterio externo, ¿por qué no obtenerla en vez de depender de la prueba? Hay dos razones; primero, puede ser que no podamos obtener la información sobre el criterio sino después de cierto tiempo y nos gustaría predecir, ahora, el estatus en que se encontrará la persona en el futuro en relación con ese criterio. Segundo, puede ser que obtener información del criterio requiere demasiado tiempo o recursos y nos gustaría usar un

método más sencillo para estimar cuál podría ser el estatus de la persona. En cualquiera de los dos casos, determinaremos si la prueba proporciona información útil acerca del probable estado de la persona en relación con el criterio externo. Consideremos primero algunos ejemplos de este tipo de validez de criterio y luego examinemos con exactitud cómo se expresa el grado de validez. El cuadro 5-7 presenta diversos ejemplos del uso de una prueba para estimar el estatus en relación con algún criterio externo. Por ejemplo, como señalamos antes, podemos usar una prueba de ingreso a la universidad para predecir el GPA al final del primer año en la universidad. Del mismo modo, podemos usar la prueba para predecir el desempeño en un trabajo de acuerdo con las valoraciones del supervisor al final de los primeros seis meses en el puesto. Quizá queramos determinar la gravedad de una depresión; podríamos tener tres clínicos que entrevistan, por separado, a un cliente durante una hora y juzgan su grado de depresión. Esto es muy caro; por ello, queremos saber qué tan bien una prueba de 15 minutos indicará el grado de la depresión. En cada uno de estos casos, tenemos un criterio externo que define lo que en realidad queremos saber. Podemos considerar la prueba que estamos validando como un potencial sustituto del criterio externo.

Cuadro 5-7. Ejemplos de criterios externos usados para establecer la validez de criterio de una prueba

Prueba	**Criterio**
Prueba de admisión a la universidad	GPA al término del primer año en la universidad
Inventario de depresión	Valoración del clínico sobre la gravedad de la depresión
Prueba de habilidades de oficina	Valoración del supervisor acerca del desempeño laboral
Prueba de pensamiento creativo	Valoración de un panel acerca de la creatividad manifestada en producciones artísticas
Escala de personalidad del vendedor	Dinero de seguros vendidos en un año

En estas situaciones, por lo general expresamos la validez de la prueba por medio de un coeficiente de correlación. Casi siempre usamos el ya familiar coeficiente de correlación de Pearson, aunque otros tipos de coeficientes también pueden usarse dependiendo de la naturaleza de las escalas que constituyen el criterio y la prueba. Cuando se usa el coeficiente de correlación de esta manera, se denomina **coeficiente de validez**. Por lo común, un coeficiente de validez es un simple coeficiente de correlación usado para expresar la validez de una prueba. De ahí que todo lo que hemos aprendido acerca de los coeficientes de correlación se pueda aplicar a los coeficientes de validez. La figura 5-3 muestra una distribución bivariada y el coeficiente de correlación resultante que expresa la validez de una prueba de ingreso a la universidad.



	M	DE
Prueba (X)	5.68	2.25
Criterio (Y)	5.20	2.37

N = 50
r = .606

Nota: En algunas ocasiones, un solo punto de la distribución puede representar más de un caso.

Figura 5-3. Distribución bivariada que ilustra la relación entre una prueba y un criterio externo.

Recordemos de nuestra revisión de las correlaciones del capítulo 4 que, una vez que conocemos la correlación entre dos variables, podemos usarla para predecir el valor de la variable *Y* a partir del valor de la variable *X*. (Usamos la palabra *predecir* para referirnos tanto a la forma predictiva como a la concurrente de la validez de criterio.) En el contexto de la validez de criterio, *Y* es el criterio externo y *X* es la prueba; por lo tanto, podemos aplicar la ecuación de regresión común:

$$Y' = bX + a$$

Fórmula 5-1

Y' = valor predicho del criterio
X = puntuación de la prueba
b = pendiente de la línea de regresión
[a] = intersección en la variable *Y*

Cuando tenemos las medias y desviaciones estándar de las variables *X* y *Y*, así como la correlación entre ellas, la fórmula más conveniente para la ecuación de la regresión es:

$$Y' = r_{XY}(DE_Y/DE_X)\ (X - M_X) + M_Y$$

Fórmula 5-2

r_{XY} = correlación entre prueba y criterio
DE_Y = desviación estándar del criterio
DE_X = desviación estándar de la prueba
X = puntuación de la prueba
M_X = media de la prueba
M_Y = media del criterio

Usar esta ecuación de regresión a menudo es desconcertante para los estudiantes. Se preguntan: si ya tenemos las puntuaciones de X y Y para determinar r_{XY}, ¿por qué necesitamos hacer una predicción de Y? La respuesta es que determinamos r_{XY} en una investigación; luego, en otra situación, cuando *no* tengamos las puntuaciones Y, podemos usar la información que obtuvimos en la investigación y nuestro conocimiento de la ecuación de regresión para predecir Y.

Recordemos también el concepto de error estándar para la ecuación de la regresión. Éste es el **error estándar de estimación** ($EEE_{Y'}$), que se expresa así:

$$EEE_{Y'} = DE_Y \sqrt{1 - r_{XY}^2}$$

Fórmula 5-3

DE_Y = desviación estándar de las puntuaciones criterio
r_{XY} = correlación (coeficiente de validez) entre el criterio (Y) y la prueba (X)

Ésta es igual a la fórmula 4-4 del capítulo 4. Recordemos que debemos distinguir entre los tres tipos de error estándar que hemos encontrado hasta aquí: el error estándar de la media usado en relación con la variabilidad del muestreo, el error estándar de medición usado con la confiabilidad y el error estándar de estimación. La comparación entre las fórmulas se puede ver en la página 95.

El error estándar de estimación *es* una desviación estándar de las puntuaciones reales del criterio alrededor de las puntuaciones predichas. De acuerdo con el supuesto de homocedasticidad y nuestro conocimiento de la curva normal, podemos aplicar esta fórmula para estimar las probabilidades de que los casos individuales estén por encima o por debajo del estatus predicho del criterio externo en ciertas cantidades. La figura 5-4 presenta el modelo para hacer tales predicciones.

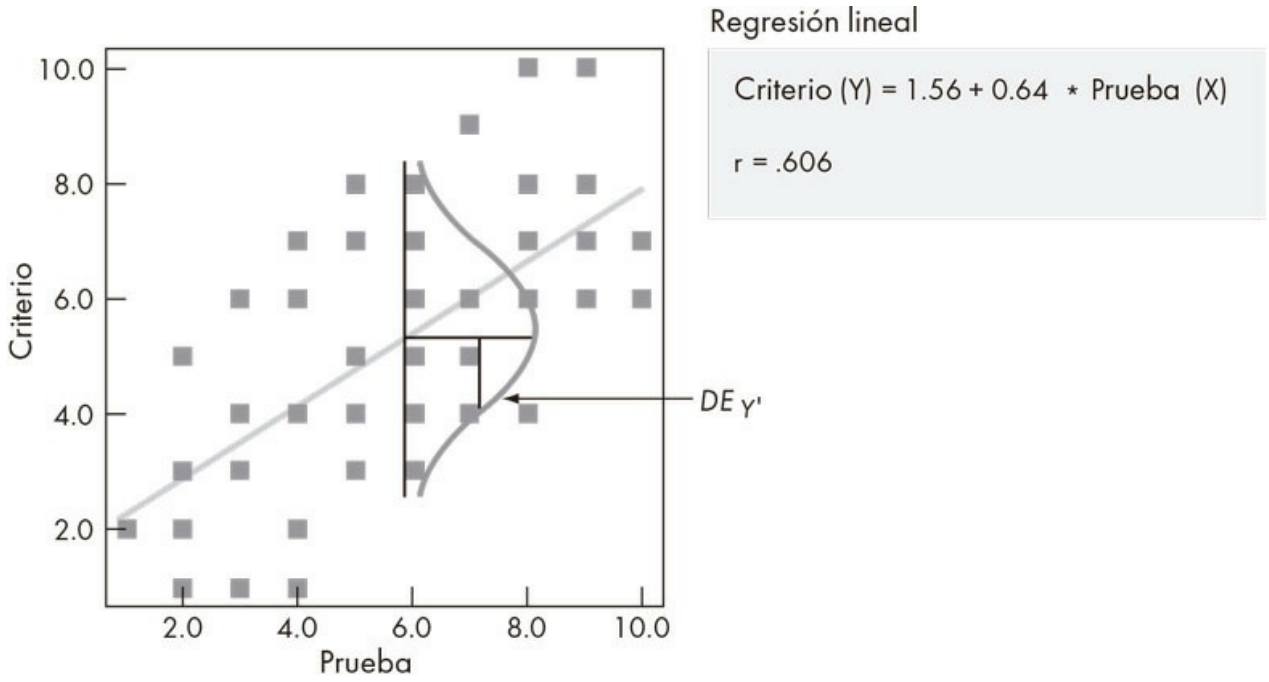


Figura 5-4. Línea de predicción y dispersión alrededor de ella.

¡Inténtalo!

En el caso de los datos de la figura 5-4, ¿cuál es el error estándar de estimación? Usa la fórmula 5-3 de la DE.

Grupos contrastados

El segundo método para demostrar la validez de criterio es el de grupos contrastados. En este caso, el criterio es la pertenencia a un grupo. Queremos demostrar que la prueba diferencia un grupo de otro. Por lo general, mientras mejor es la diferenciación entre grupos, más válida es la prueba. Suponemos que la pertenencia grupal es una buena definición del criterio. Podemos ilustrar este método con unos pocos ejemplos.

En el primer ejemplo, el grupo A consta de 50 individuos diagnosticados con esquizofrenia; el diagnóstico se basa en entrevistas exhaustivas llevadas a cabo por tres clínicos independientes, por lo que podemos confiar en él. El grupo B consta de 50 individuos sin un historial con problemas psicológicos importantes y, según se sabe, su funcionamiento es normal en el ambiente familiar y laboral. Aplicamos un inventario de personalidad a los 100 individuos para mostrar que la prueba distingue finamente entre ambos grupos.

En el segundo ejemplo, el grupo A consta de 35 individuos que han terminado exitosamente un curso de programación de computadora, mientras que el grupo B consta de 35 individuos que no tuvieron un buen desempeño en todo el curso. Habíamos

aplicado una prueba de aptitud de programación de computadoras a los 70 individuos antes de empezar el curso. Queremos establecer que la prueba de aptitud puede distinguir con claridad entre los exitosos y los no exitosos del curso.

Al ver los resultados de un estudio sobre grupos contrastados en relación con la validez, es importante considerar el grado de separación entre los grupos. No es suficiente limitarse a informar que hubo una "diferencia estadísticamente significativa" entre los grupos, como se hace a menudo en los manuales. Si el estudio incluye un gran número de casos, no es difícil obtener una diferencia significativa entre los grupos; lo importante es si la prueba distingue entre ellos hasta el grado de ser útil en la práctica. La significancia estadística es una condición necesaria, pero no suficiente para ser útil con fines prácticos.

Consideremos el grado de diferenciación entre grupos en los dos ejemplos de la figura 5-5. En el ejemplo A, aunque hay una diferencia significativa en las puntuaciones de las medias del criterio y los grupos contrastados, hay una superposición casi completa en las distribuciones de las puntuaciones. En el caso de casi cualquier puntuación de la prueba, es difícil conjeturar si el examinado es más como el grupo criterio o como el grupo de contraste. En el ejemplo B, hay una buena diferenciación entre los grupos; un examinado con una puntuación dentro del rango Q tiene una puntuación como el grupo de contraste, no como el grupo criterio. Un examinado con una puntuación dentro del rango S tiene una puntuación como el grupo criterio, no como el grupo de contraste. Sólo en el rango R, la información de la prueba es inútil.

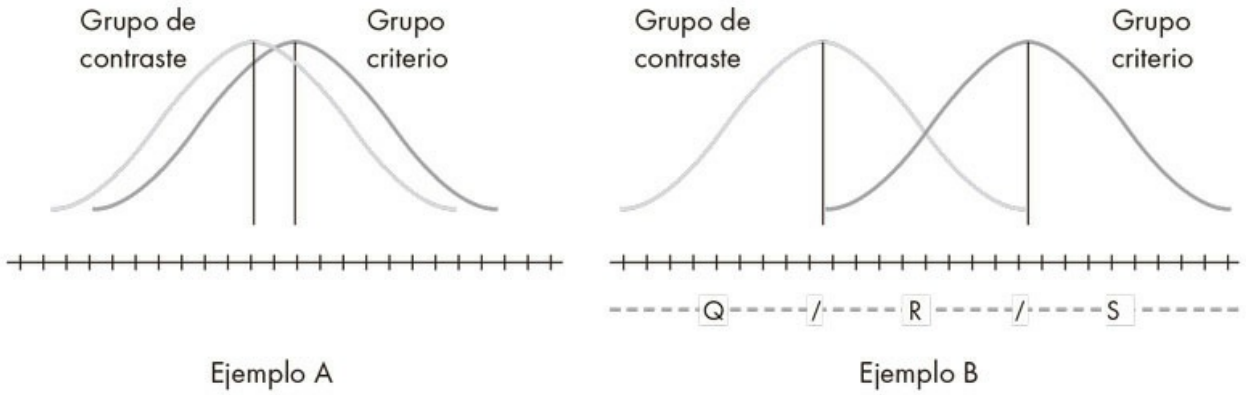


Figura 5-5. Ejemplos de una diferenciación pobre y otra buena al usar el método de grupos contrastados.

Los ejemplos de la figura 5-5, por supuesto, hacen pensar en la noción de **tamaño del efecto** de estadística básica. Sería útil aplicar esta noción a los estudios con grupos contrastados sobre la validez de criterio. Desafortunadamente, las medidas del tamaño del efecto casi no se emplean para ese propósito; sin embargo, encontraremos algunas nociones análogas en la sección sobre teoría de la decisión más adelante en este capítulo.

Como señalamos en el capítulo 3, los resultados de los estudios con grupos

contrastados constituyen la base de algunas afirmaciones en los informes interpretativos sobre las pruebas. Por ejemplo, "las personas con puntuaciones como la de Juan a menudo muestran dificultades en las relaciones interpersonales" es una afirmación que probablemente surgió de un estudio que muestra diferentes distribuciones de las puntuaciones de personas con y sin relaciones interpersonales difíciles.[1] Una afirmación como "la puntuación de Luis en la escala A sugiere un pronóstico favorable en una terapia de corto plazo" probablemente se basa en un estudio de diferencias en la escala A en personas que se beneficiaron o no de una terapia de corto plazo. Es evidente que la validez de estas afirmaciones depende, en parte, de qué tan bien la prueba diferencia los grupos.

El lector perspicaz notará que el enfoque de los grupos contrastados puede convertirse en una forma del enfoque del criterio externo y factible con el simple hecho de asignar valores de 0 y 1 a la pertenencia a los grupos. Una persona con inclinación a la estadística puede hacer con facilidad la conversión; sin embargo, en la práctica, los dos enfoques suelen tratarse como casos distintos.

Correlaciones con otras pruebas

Un tercer método para establecer la validez de criterio es mostrar la correlación entre la prueba que se desea validar y alguna otra que se sabe o se supone que es una medida válida del constructo pertinente. Por simplicidad, nos referiremos a la prueba que se desea validar como "nueva". En esta aplicación, la otra prueba se convierte en el criterio, análogo al criterio externo que tratamos antes. Al encontrarse por primera vez con este método, tendemos a preguntar: si se sabe o se supone que la otra prueba es válida, ¿por qué no usarla en vez de la nueva prueba? Hay varias razones por las que desearíamos establecer la validez de la nueva prueba; ésta puede ser más corta o menos costosa que la prueba criterio. Por ejemplo, podríamos tener una prueba de inteligencia aplicable en 15 minutos que queremos validar frente a la Escala Wechsler de Inteligencia para Niños, cuya aplicación toma alrededor de una hora. La nueva prueba puede tener mejores normas o procedimientos de calificación más eficientes; por ejemplo, podemos querer mostrar la correlación entre una nueva edición de una prueba de depresión –ahora calificada por computadora, reactivos actualizados y nuevas normas nacionales– y la edición anterior de la prueba. ¿Por qué? Porque tenemos 20 años de investigación sobre la edición anterior, pues se trata de una medida respetable que ha resistido la prueba del tiempo y muy arraigada en la literatura de la investigación sobre la depresión. Esperamos que nuestra nueva edición tenga una correlación alta con la anterior. Por cualquiera de estas razones u otras similares, podemos querer establecer la validez de la nueva prueba en vez de depender de la prueba criterio.

Usar otra prueba para establecer la validez de criterio es sencillo y un método muy empleado (Hogan & Agnello, 2004). La correlación (casi siempre la de Pearson) entre la nueva prueba y la prueba criterio expresa esa validez. Así, la metodología es la misma que se describió antes en el caso del criterio externo y factible.

Al considerar la validez de una prueba, debemos estar alerta de no confundir las palabras con la realidad. Hace muchos años, Kelley (1927) describió lo que llamó *f*alacia del retintín y falacia del tintineo. En términos sencillos, la falacia del retintín es la idea de que usar las mismas o similares palabras para nombrar dos cosas significa que en verdad son lo mismo. Aplicada a las pruebas, esta falacia implica creer que el *Wisconsin Intelligence Test* [Prueba Wisconsin de Inteligencia] y el *Scranton Intelligence Test* [Prueba Scranton de Inteligencia] miden el mismo rasgo sólo porque ambos incluyen la palabra "*intelligence*" en su nombre. Estas pruebas pueden no medir el mismo rasgo; por ejemplo, pueden tener una correlación de sólo .45, lo que sugeriría que al menos en parte miden rasgos diferentes. La falacia del tintineo es la idea de que dos cosas son en verdad diferentes, porque se usan palabras distintas para nombrarlas. Aplicada a las pruebas, esta falacia implica creer que el *Non-Verbal Test of Intelligence* [Prueba de Inteligencia No Verbal] y el *Test of Verbal Fluency* [Prueba de Fluidez Verbal] miden rasgos diferentes porque tienen palabras distintas en sus nombres. Estas dos pruebas pueden o no medir rasgos distintos; por ejemplo, su correlación podría ser de .95, lo que sugeriría con fuerza que ambas miden el mismo rasgo. Para protegernos de estas falacias, es necesaria la evidencia empírica; la información sobre las correlaciones entre las pruebas es en especial pertinente, aunque no es la única.

El índice más común para informar la validez de criterio es el coeficiente de correlación, cuyo grado puede representarse por medio de distribuciones bivariadas, como las que se presentan en las figuras 5-3 y 5-4. Una aplicación especial de esta disposición es el *cuadro de* expectativas, que tiene una estructura muy similar a la de la gráfica bivariada. Las entradas de cada fila en el cuadro de expectativas son porcentajes de los casos en ella. Así, entradas y combinaciones de entradas se pueden traducir con facilidad a probabilidades. En la era previa a la computadora, los cuadros de expectativas facilitaron la interpretación de los datos de la validez de criterio. La disponibilidad de las predicciones generadas por computadora, usando los métodos antes descritos en este capítulo, volvió obsoletos los cuadros de expectativas. Sin embargo, los manuales de pruebas de tiempo atrás aún contienen dichos cuadros.

Consideraciones especiales para interpretar la validez de criterio

A primera vista, la validez de criterio parece clara y sencilla, y de hecho lo es de varias maneras. Sin embargo, bajo esa apariencia de sencillez se esconden numerosos problemas y cuestiones que merecen especial atención. Ahora nos ocuparemos de esas cuestiones especiales.

Condiciones que afectan el coeficiente de correlación

Al revisar el coeficiente de correlación (*r*) en el capítulo 4, señalamos varias condiciones que afectan su magnitud. Ya que el coeficiente de validez es simplemente un tipo de

coeficiente de correlación, puede ser pertinente considerar todas estas condiciones en relación con la validez de criterio: en particular, linealidad, homogeneidad del grupo y heterocedasticidad son cuestiones importantes.

Si la relación entre prueba y criterio es *no lineal*, la correlación de Pearson subestimará su verdadera magnitud. Al usar el coeficiente de correlación para expresar la validez de criterio, siempre debemos examinar la distribución bivariada (diagrama de dispersión) de las dos variables. La no linealidad en la relación *no* es un problema común cuando estudiamos la validez de las pruebas, pues las correlaciones de las pruebas con otras variables no son, por lo común, lo suficientemente fuertes para poner de manifiesto tendencias claramente no lineales. No obstante, es fácil revisar la distribución bivariada para determinar la presencia de una tendencia no lineal. Lo único que hay que hacer es examinar la gráfica bivariada (dispersograma), que por lo general se puede crear en SPSS, SAS, Excel o programas similares, y determinar si existe una tendencia no lineal.

La diferencia en la heterogeneidad de grupo es un problema común cuando interpretamos los coeficientes de validez. Un estudio sobre la validez puede llevarse a cabo con un grupo muy heterogéneo, lo que producirá un coeficiente de validez relativamente alto, cuando queremos aplicar el resultado a un grupo mucho más homogéneo. Por ejemplo, la validez de una prueba de admisión a la universidad para predecir el GPA de alumnos de primer año puede establecerse en un estudio en varios campus, que incluya un amplio rango de capacidades. Queremos usar la prueba en un solo campus, donde el rango de capacidades es mucho más limitado. Casi con toda certeza la validez será menor en nuestro único campus. Por el contrario, podemos llevar a cabo un estudio en un sólo campus con un rango limitado de talento; con seguridad, la prueba tendrá una mayor validez predictiva en las escuelas con un rango más amplio de capacidades. En el capítulo 4, presentamos las fórmulas para hacer los ajustes apropiados respecto de las diferencias en la heterogeneidad grupal. Estas fórmulas se usan de manera rutinaria en el estudio de la validez de criterio.

La homocedasticidad, descrita en el capítulo 4, se refiere al supuesto de que los puntos de los datos están dispersos de un modo aproximadamente igual alrededor de la línea de predicción a lo largo de todo el rango. Véase la figura 5-4. Por lo general, esto *no* es un problema cuando examinamos la validez de la prueba. Las correlaciones entre las puntuaciones de la prueba y otros criterios a menudo no son lo suficientemente altas para tener que preocuparnos por ello. Sin embargo, al igual que con la no linealidad, es fácil revisar el dispersograma para determinar si hay un problema al respecto.

Relación entre confiabilidad y validez «117-118a «117-118b

La validez de una prueba depende en parte de su confiabilidad y, en parte, de la confiabilidad del criterio. Así, una confiabilidad limitada, sea de la prueba o del criterio, limitará la validez de criterio. Estas relaciones entre confiabilidad y validez suelen abordarse en el contexto de la validez de criterio, una costumbre que aquí también adoptamos. Sin embargo, las nociones fundamentales se extienden más ampliamente a

todo tipo de validez. Los conceptos que revisamos en esta sección están entre los más importantes de toda la teoría psicométrica.

Primero, expresamos algunas relaciones entre confiabilidad (tanto de la prueba como del criterio) y la validez en una forma narrativa. Después, examinaremos las relaciones de una manera más formal con las fórmulas que las expresan. Si una prueba no tiene ninguna confiabilidad –las puntuaciones de la prueba son sólo error aleatorio–, tampoco puede tener validez; sin embargo, una prueba puede ser por completo confiable y, aun así, carecer de validez; es decir, la prueba es confiable midiendo algo diferente de lo que queremos medir. Si el criterio no tiene confiabilidad –su estatus es sólo error aleatorio–, la prueba no puede tener validez respecto del criterio, aunque la prueba sea por completo confiable. Hemos formulado las últimas afirmaciones en términos de extremos: nada de confiabilidad y por completo confiable. Desde luego, en la práctica, lo usual es encontrar casos menos extremos, pues las pruebas y los criterios suelen tener algún grado de confiabilidad. ¿Qué hacemos en estos casos intermedios?

Por fortuna, existen fórmulas que expresan el efecto de una confiabilidad limitada sobre la validez de criterio y que, también por fortuna, son sencillas aunque no obvias para el sentido común. Antes de citar las fórmulas pertinentes, presentaremos los términos especializados que usamos para hablar de este tema. **Atenuación** es un término técnico que se refiere al límite impuesto a la validez debido a la confiabilidad imperfecta; significa sencillamente "disminución" o "reducción". A partir del coeficiente de validez obtenido, podemos calcular el coeficiente de validez *desatenuado*, que también se denomina coeficiente de validez corregido por falta de confiabilidad. Podemos corregir o desatenuar el coeficiente de validez por falta de confiabilidad tanto de la prueba como del criterio. Estas correcciones proporcionan el coeficiente de validez estimada si la confiabilidad (de la prueba, del criterio o de ambos) es perfecta, es decir, +1.00. Éstos son los símbolos que usamos en las fórmulas de corrección:

Y = criterio
X = prueba
r_{XY} = correlación entre prueba y criterio (coeficiente de validez)
r_{XX} = confiabilidad de la prueba
r_{YY} = confiabilidad del criterio

Y éstas son las fórmulas apropiadas. Utilizamos el símbolo de prima (') en X, Y o ambas para indicar que hemos corregido la correlación por falta de confiabilidad en la(s) variable(s).

$$r_{X'Y} = \frac{r_{XY}}{\sqrt{r_{XX}}}$$

Fórmula 5-4

La fórmula 5-4 proporciona el coeficiente de validez corregido por falta de confiabilidad en la *prueba* (*X*). Una modificación a esta fórmula origina la generalización concisa de que el coeficiente de validez no puede exceder la raíz cuadrada de la confiabilidad de la prueba (*X*). Quizá es más importante sólo recordar que la validez de una prueba está limitada por su confiabilidad.

$$r_{XY'} = \frac{r_{XY}}{\sqrt{r_{YY}}}$$

Fórmula 5-5

La fórmula 5-5 proporciona el coeficiente de validez corregido por falta de confiabilidad en el *criterio* (*Y*).

$$r_{X'Y} = \frac{r_{XY}}{\sqrt{r_{XX}\ r_{YY}}}$$

Fórmula 5-6

La fórmula 5-6 proporciona el coeficiente de validez corregido por falta de confiabilidad tanto en la prueba como en el criterio. En Gulliksen (1950), Lord y Novick (1968) y Nunnally y Bernstein (1994) se pueden encontrar más detalles sobre estas fórmulas.

Consideremos este ejemplo. La correlación entre una prueba (*X*) diseñada para predecir el éxito en un trabajo, el cual es definido por la valoración del supervisor del desempeño (*Y*, el criterio), es .60. La confiabilidad de la prueba es .75. Si la prueba tuviera una confiabilidad perfecta, la correlación entre prueba y criterio (el coeficiente de validez) sería

$$.60/\sqrt{.75} = .75$$

Supongamos que la confiabilidad de la valoración del supervisor es .65. La corrección por falta de confiabilidad tanto en la prueba como en el criterio produce un coeficiente de validez de

$$.60/\sqrt{.75x.65} = .86$$

Así, el coeficiente de validez (.60), que es moderado, está limitado considerablemente por la confiabilidad imperfecta de la prueba y el criterio.

En la mayoría de las aplicaciones prácticas de estos procedimientos, corregimos sólo por falta de confiabilidad en la prueba. Suponemos que la confiabilidad del criterio es irreprochable o, de modo más realista, que no hay nada que hacer al respecto. Sin embargo, a veces es útil aplicar la corrección también al criterio. Es importante hacer hincapié en que aplicar estas correcciones no cambia en realidad el coeficiente de validez determinado en un estudio específico. No obstante, las correcciones nos ayudan a pensar en los efectos de la confiabilidad imperfecta sobre el coeficiente de validez.

¡Inténtalo!

Aplica la corrección por falta de confiabilidad (sólo de la prueba) a estos datos:
r_{XY} = .40 r_{XY} = .70 $r_{X'Y}$ =

La corrección por falta de confiabilidad suele aplicarse para llevar la prueba a un nivel de confiabilidad perfecta (1.00). Aunque este procedimiento es útil para propósitos teóricos, es muy poco realista. Es más realista fijar la confiabilidad en una cifra como .85 o .90, lo cual puede hacerse incluyendo una cifra más realista como multiplicador en el denominador de las fórmulas citadas antes. Por ejemplo, la primera fórmula puede escribirse como

$$r_{X'Y} = \frac{r_{XY}}{\sqrt{.90(r_{XX})}}$$

Fórmula 5-7

Esto nos dará un coeficiente de validez estimada (r_{XY}) con el supuesto de que la confiabilidad de la prueba (r_{XX}) es elevada a .90.

Validez del criterio

Al discutir la validez de criterio, tendemos a centrar la atención en la prueba. ¿Qué tan bien la prueba predice o se correlaciona con el criterio? De hecho, la prueba debe ser el centro de atención, porque tratamos de evaluar su validez. Sin embargo, desde otra perspectiva, necesitamos examinar la validez del criterio, en especial la definición operacional del criterio; ¿es apropiada?

Consideremos algunos ejemplos. Queremos que una prueba de admisión a la universidad prediga el "éxito académico". Utilizamos el GPA como definición operacional del éxito en la universidad. ¿Qué tan buena es esta definición operacional? El GPA de los

alumnos de primer año es sólo una posible definición del éxito en la universidad. Otra posibilidad es el GPA tras la graduación. Una posibilidad más es la participación activa en actividades extracurriculares o una calificación compuesta por el GPA y la participación extracurricular. ¿Qué hay del éxito como vendedor? El volumen total de dólares vendidos podría ser una buena definición de éxito, pero quizá no sea la mejor. Algunos vendedores pueden ser asignados a áreas del mercado que por supuesto tienen un alto volumen. Quizá el número de nuevas cuentas adquiridas sería una mejor definición de éxito, o la valoración del jefe de vendedores podría servir como definición de éxito. Obviamente, podríamos dar múltiples ejemplos de distintas maneras de definir cualquier criterio que pudiera usarse para la validación de una prueba. Lo importante aquí es que al considerar la validez de criterio de una prueba, también necesitamos pensar en la validez de la definición operacional del criterio.

Contaminación del criterio

Cuando tratamos de establecer la validez de una prueba correlacionándola con un criterio externo, la **contaminación del criterio** se refiere a una situación en la que el desempeño en la prueba influye en el estatus del criterio. Un ejemplo pondrá en claro el concepto. Con una muestra de 50 casos, intentamos establecer la validez del *Cleveland Depression Scale* (CDS [Escala Cleveland de Depresión]) mostrando que tiene una correlación alta con las valoraciones de la depresión realizadas por tres clínicos. Éstos tienen acceso a las puntuaciones, por lo que basan su valoración, al menos en parte, en ellas. Esto conduce a inflar la correlación entre la prueba y el criterio. También es posible que la influencia sea en la dirección contraria, es decir, que la correlación disminuya; por ejemplo, si los clínicos desprecian el CDS, podrían estar en desacuerdo con él de modo deliberado. Sin embargo, esto es poco probable; la contaminación del criterio, por lo general, lleva a aumentar la correlación entre prueba y criterio.

Cuando se lleva a cabo un estudio de validez de criterio, es importante que el diseño evite la contaminación del criterio. Cuando se revisa uno de estos estudios, debemos estar alerta para detectar la posible presencia de la contaminación del criterio. No existen métodos analíticos ni fórmulas que estimen el efecto de esta contaminación.

Validez convergente y discriminante

Dos conceptos útiles para pensar acerca de la validez de criterio son la validez convergente y la validez discriminante. La **validez convergente** se refiere a una correlación relativamente alta entre la prueba y algún criterio pensado para medir el mismo constructo que la prueba; por ejemplo, para demostrar la validez de una prueba de depresión, podemos querer mostrar que ésta tiene una correlación alta con otra prueba reconocida como una buena medida de depresión. En contraste, podemos querer mostrar que nuestra prueba de depresión *no* es una simple medida de inadaptación general, por lo

que queremos mostrar que *no* tiene una correlación alta con constructos como ansiedad o estrés. Ésta es la **validez discriminante**, la que muestra que una prueba tiene una correlación relativamente baja con constructos diferentes al que se pretende medir con ella.

> Las relaciones entre las puntuaciones de la prueba y otras medidas del mismo constructo, o similares, proporcionan evidencia convergente, mientras que las relaciones entre las puntuaciones de la prueba y medidas de constructos manifiestamente diferentes proporcionan evidencia discriminante.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Los conceptos de validez convergente y discriminante se usan mucho en el campo de la medición de la personalidad, pero se usan poco en el de las pruebas de capacidad y aprovechamiento en la práctica, aunque estos conceptos, sin duda, tienen aplicaciones potenciales en estas áreas. Aquí presentamos un ejemplo de cómo se usan estos conceptos en el campo de la personalidad; supongamos que intentamos establecer la validez del *Scranton Test of Anxiety* (STA) tratando de mostrar que su correlación con otras medidas de ansiedad es alta y que con medidas de depresión *no*. Aplicamos el STA junto con el *Taylor Manifest Anxiety Scale* (TMAS [Escala de Ansiedad Manifiesta de Taylor]) y el *Beck Depression Inventory* (BDI) suponiendo que son medidas razonablemente válidas de ansiedad y depresión, respectivamente. Un resultado favorable sería encontrar correlaciones de .75 entre STA y TMAS (validez convergente) y de .20 entre STA y BDI (validez discriminante). Pero si la correlación entre STA y BDI fuera de .75, concluiríamos que el STA no discrimina entre ansiedad y depresión. Este tipo de análisis y razonamiento es muy común en las discusiones sobre la validez de las pruebas de personalidad. En el cuadro 5-8 se encuentran afirmaciones en las que se emplean estos conceptos.

Cuadro 5-8. Afirmaciones muestra de la validez convergente y discriminante provenientes de manuales de pruebas

"Es de especial importancia el hallazgo de que el BDI-II [*Beck Depression Inventory*-II] tuvo una correlación positiva más alta (r = .71) con el *Hamilton Psychiatric Rating Scale for Depression* (HRSD [Escala Hamilton de Valoración Psiquiátrica de Depresión])... que con el *Hamilton Rating Scale for Anxiety* (HRSA [Escala Hamilton de Valoración de Ansiedad]) (r = .47)... Estos hallazgos indican una validez discriminante robusta entre depresión y ansiedad." (Beck, Steer, & Brown, 1996, p. 28)
"Las correlaciones de las escalas STAI [*State-Trait Anxiety Inventory* {Inventario de Ansiedad Rasgo-Estado}] y otras medidas de personalidad proporcionan evidencia de la validez convergente y discriminante del STAI. En general, se esperarían correlaciones mayores con las medidas de perturbación emocional y psicopatología, y correlaciones menores con constructos no relacionados." (Spielberger, 1983, p. 35)
"Un estudio... comparó el Piers-Harris [Children's Self Concept Scale {Escala Piers-Harris de Autoconcepto Infantil}] con el Cooper-Smith Self-Esteem Inventory [Inventario Cooper-Smith de Autoestima]... Las dos medidas tuvieron una correlación de r = .78, que establece la validez

convergente. La validez discriminante se evaluó correlacionando las puntuaciones del autoconcepto con variables que representan el aprovechamiento académico, el estatus socioeconómico, ubicación en educación especial, origen étnico, grado, género y edad. Los coeficientes de correlación múltiple con estas variables conceptualmente distintas no superaron el .25, lo que constituye evidencia de la validez discriminante." (Piers & Herzberg, 2002, p. 66)
"El hecho de que las escalas NEO PI-R se correlacionen con medidas alternativas de constructos similares es una evidencia de su validez convergente... La validez discriminante se observa contrastando sus correlaciones con aspectos diferentes dentro del mismo dominio." (McCrae & Costa, 2010, p. 74)

Matriz multirrasgo-multimétodo

Una aplicación especial de los conceptos de validez convergente y discriminante es la matriz multirrasgo-multimétodo. En un artículo clásico, Campbell y Fiske (1954) recomendaron el uso de esta matriz para analizar la validez convergente y divergente de varias pruebas. La matriz es justo una matriz de correlaciones, donde las variables incluyen pruebas que pretenden medir diferentes rasgos –por eso es multirrasgo– por medio de distintos métodos –por eso es multimétodo. Los distintos rasgos podrían ser ansiedad y depresión, como en el ejemplo anterior, mientras que los distintos métodos podrían incluir cuestionarios de autorreporte, técnicas proyectivas y valoraciones basadas en entrevistas clínicas. El propósito esencial del **análisis multirrasgo-multimétodo** es demostrar que las correlaciones dentro de un rasgo utilizando distintos métodos son más altas que las correlaciones dentro de un método con distintos rasgos y, desde luego, que las correlaciones que combinan distintos rasgos y métodos. El cuadro 5-9 presenta un esquema para comprender la matriz multirrasgo-multimétodo. En este ejemplo, intentamos medir depresión y ansiedad (dos rasgos supuestamente distintos). De cada rasgo tenemos una medida proyectiva, digamos una puntuación de la prueba de manchas de tinta Rorschach, y la puntuación de un inventario de autorreporte, digamos el MMPI-2. Llamemos los rasgos D y A, y los métodos 1 y 2. En el cuadro 5-9, las correlaciones (*r*) en diagonal son los coeficientes de confiabilidad. Las otras entradas también son correlaciones codificadas en términos de nuestras expectativas acerca de sus niveles. "CA" significa que esperamos encontrar correlaciones altas; por ejemplo, queremos obtener una correlación alta entre las dos medidas de depresión a pesar de que son medidas derivadas de métodos diferentes, en este caso, el Rorschach y el MMPI-2. "CB" y "CMB" significan que esperamos encontrar una correlación baja y muy baja, respectivamente. Queremos encontrar una correlación baja entre depresión y ansiedad aunque ambas sean medidas por el Rorschach. Desde luego, esperamos encontrar una correlación muy baja entre depresión, medida con el MMPI-2, y ansiedad, medida con el Rorschach.

Cuadro 5-9. Ejemplo sencillo de una matriz multirrasgo-multimétodo

	D-1	**D-2**	**A-1**	**A-2**

D-1	*r*			
D-2	CA	*r*		
A-1	CB	CMB	*r*	
A-2	CMB	CB	CA	*r*

El cuadro 5-10 presenta datos ilustrativos de la matriz multirrasgo-multimétodo de nuestro ejemplo. En él, las correlaciones se reflejan de manera favorable en las pruebas, es decir, muestran una validez convergente y divergente apropiada.

Cuadro 5-10. Datos ilustrativos de una matriz multirrasgo-multimétodo

	D-1	**D-2**	**A-1**	**A-2**
D-1	.84			
D-2	.75	.87		
A-1	.32	.17	.79	
A-2	.09	.49	.65	.81

Queremos hacer hincapié en que éste es un ejemplo muy sencillo empleado con fines didácticos. Campbell y Fiske (1954) emplearon ejemplos que incluían tres rasgos y tres métodos o más; la matriz de correlaciones puede volverse muy grande con rapidez. Este método se cita mucho en la literatura psicométrica; sin embargo, en la práctica *no* se usa tanto. En un artículo posterior al de 1954, Fiske y Campbell (1992) lamentaron que, mientras que su artículo de 1954 fue citado miles de veces, "aún nos queda por ver una matriz en verdad buena" (p. 393). No obstante, este enfoque nos ayuda a pensar con mayor claridad acerca de nuestros métodos para validar pruebas.

Combinación de información de diferentes pruebas

Hasta ahora, nos hemos referido a la validez de criterio como la relación entre una sola prueba y un criterio; sin embargo, en algunos contextos queremos usar varias pruebas para predecir el estatus de un criterio. El método usual para tratar con esta situación es la **correlación múltiple**, técnica para expresar la relación ente una variable (el criterio) y la combinación óptima de dos o más variables (en este caso, varias pruebas). Por ejemplo, podemos querer predecir el GPA de un estudiante de primer año a partir de la combinación de una prueba de admisión, su rango en el bachillerato y una prueba de motivación académica. El truco es definir los pesos óptimos de las variables para maximizar la correlación entre el criterio y la combinación de pruebas. Estos pesos dependen no sólo de las correlaciones de las pruebas con el criterio, sino también de las relaciones entre ellas.

Hay dos propósitos principales de los procedimientos de correlación múltiple. El primero es muy práctico; se trata de obtener la mejor predicción posible de una variable

dependiente, como el éxito en un trabajo o en el desempeño académico, a partir de otras variables y de la manera más económica posible, es decir, sin incluir ninguna variable más que las necesarias. El segundo propósito es comprender a nivel teórico qué variables contribuyen efectivamente a la predicción de una variable dependiente y qué variables son superfluas.

Hay dos productos finales de los procedimientos de correlación múltiple. El primero es un coeficiente de correlación múltiple, representado mediante *R* (mayúscula). *R* se escribe acompañada de subíndices para indicar lo que se predice y a partir de qué se predice; por ejemplo, si la variable 1 se predice a partir de las variables 2, 3 y 4, escribimos $R_{1.234}$. Esta *R* se interpreta de la misma manera que la *r* de Pearson, a la que ahora llamaremos coeficiente de correlación de orden cero.

El segundo producto de los procedimientos de correlación múltiple son los pesos asignados a las variables predictoras. Éstas tienen dos formas: las *b* y las [β] (beta). Las "*b*" se aplican a las puntuaciones naturales, mientras que las "[β]", a las puntuaciones "estandarizadas", es decir, puntuaciones z. La ecuación que muestra una predicción a partir de una correlación múltiple, llamada **ecuación de regresión múltiple**, es como ésta si incluye tres predictores:

Forma con puntuaciones naturales:

$$Y' = b_1X_1 + b_2X_2 + b_3X_3 + c$$

Fórmula 5-8

Forma con puntuaciones z:

$$z_{y'} = [\beta]_1 z_1 + [\beta]_2 z_2 + [\beta]_3 z_3$$

Fórmula 5-9

Podemos notar la diferencia entre las *b* y las [β]; las primeras sólo nos dicen cuánto peso dar a cada variable de puntuación natural y compensan las diferencias en las escalas empleadas en las puntuaciones naturales. Las variables con números "grandes" por lo general obtienen pesos pequeños, y a la inversa, variables con números "pequeños" obtienen pesos grandes. En la forma con puntuación z, todas las variables tienen $M = 0$ y $DE = 1$, por lo que los pesos beta pueden compararse directamente, pues indican de manera inmediata qué variables reciben la mayor parte del peso.

También usamos *R* cuadrada (R^2), es decir, el porcentaje de varianza en *Y* explicado por la varianza de los predictores o que se superpone a ella. Esto nos lleva a una manera interesante y útil en términos didácticos de interpretar las contribuciones de diferentes variables en *R*.

Consideremos los ejemplos de la figura 5-6. El marco "1" representa el criterio que tratamos de predecir; el grado en que los marcos (2, 3 y 4) se superponen al marco 1 es proporcional a sus correlaciones respectivas con el criterio. Es decir, el grado de

superposición con el criterio corresponde a r^2 de cada una de las otras variables. De manera similar, el grado de superposición entre las cajas 2, 3 y 4 es proporcional a sus respectivas intercorrelaciones. En estos ejemplos, queremos usar las pruebas 2, 3 y 4 para predecir el criterio (1).

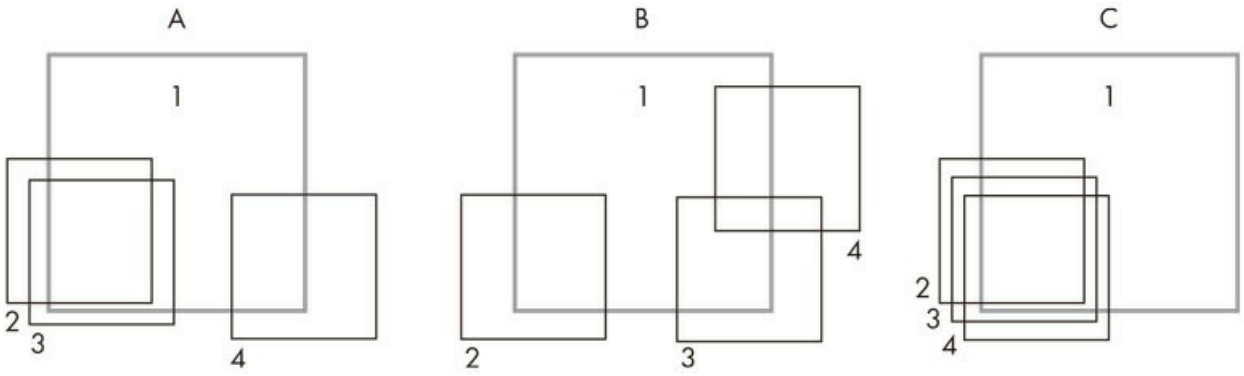


Figura 5-6. Ilustración de las posibilidades de la regresión múltiple.

Ahora consideremos el ejemplo A. Las pruebas 2 y 3 muestran una superposición considerable con el criterio; sin embargo, entre sí tienen una correlación alta. Después de que introducimos una de ellas en la fórmula para predecir el criterio (es decir, en la ecuación de regresión múltiple), la otra añade poca información nueva. Supongamos que la prueba 2 se incluye primero en la ecuación; ésta tendrá el mayor peso ([β]). La prueba 4 tendrá el siguiente peso de mayor magnitud aunque la prueba 3 tenga una correlación más alta con el criterio que la prueba 4, pues esta última agrega más información nueva o única después de que la prueba 2 ya está en la ecuación. Podemos notar que la prueba 4 no está correlacionada con las pruebas 2 y 3 (y no se superpone a ellas).

En el ejemplo C, las pruebas 2, 3 y 4 tienen casi el mismo grado de correlación (superposición) con el criterio. Además, las tres pruebas tienen correlaciones altas entre sí, como lo indica el grado en que se superponen una a otra. Después de que introducimos una de ellas en la ecuación, las otras dos agregan poca información nueva. Desde un punto de vista práctico, no valdría la pena aplicar estas tres pruebas con el fin de predecir el criterio. Por ejemplo, para predecir el GPA de un estudiante de primer año, no sería útil aplicar tres pruebas separadas de capacidad verbal general aunque las tres tengan una correlación considerable con el GPA.

Las cuestiones importantes de la metodología de correlación múltiple son: 1) el orden en que las variables se introducen en la ecuación, 2) la superposición entre los predictores y 3) cuando nuevas variables no aportan ningún poder predictivo. Así, los procedimientos de correlación múltiple pueden mostrar que ciertas variables no son predictores valiosos una vez que se han tomado en cuenta otros predictores.

La correlación múltiple es una técnica crucial para determinar la **validez incremental**, que se refiere a cuánta información nueva y única aporta una prueba (u otra fuente de

información) a la información existente. La noción general de validez incremental es importante muy aparte de la correlación múltiple. Siempre tratamos de determinar cuánta información nueva ofrece una prueba o un procedimiento, cuán difícil y costoso es obtener información nueva y si ésta vale el esfuerzo y costo extras. En algunas circunstancias, podemos estar interesados en tener una prueba con validez buena, pero no necesitamos usar la prueba, porque ya tenemos información buena acerca del rasgo de interés. En otras circunstancias, podemos no tener prácticamente ninguna información sobre el rasgo de interés, por lo que nos alegra poder usar una prueba que tenga sólo una validez modesta, pues al menos nos da alguna información útil. En Hunsley y Haynes (2003) podemos encontrar aplicaciones prácticas de la noción de validez incremental en contextos clínicos.

Los procedimientos de correlación y regresión múltiple proporcionan los detalles matemáticos de los conceptos que hemos ilustrado en la figura 5-6. Podemos entender los conceptos generales implicados sin un íntimo conocimiento de los procedimientos matemáticos; esto es suficiente para nuestros propósitos. Los procedimientos de regresión múltiple permiten diferentes modos de añadir variables en la ecuación, tema que va más allá de nuestros objetivos en este libro.

La correlación múltiple es la técnica estadística multivariada que se usa con mayor frecuencia cuando se combina información de distintas pruebas. Sin embargo, no es la única técnica multivariada para este propósito, pues existen otras, como las funciones discriminantes, las correlaciones canónicas o los modelos de ecuaciones estructurales que están más allá del alcance de este libro. En Tabachnick y Fidell (2007) se puede encontrar más información sobre estas técnicas.

Validación cruzada y encogimiento de la validez

Si dejamos que el azar opere en un número suficiente de eventos, de seguro observaremos algunos resultados inusuales. Eso es lo que sucede con los eventos climáticos, con una moneda lanzada al aire y con una correlación múltiple. Si introducimos suficientes variables en la ecuación (fórmula 5-8), algunas de ellas serán "significativas" o tendrán pesos extrañamente grandes (o pequeños), pero sólo por azar. Una práctica deseable es el uso de la **validación cruzada**, que se refiere a determinar la ecuación (y R) en una muestra, y luego aplicar la ecuación en una nueva muestra para ver qué R emerge. La pérdida de validez (es decir, reducción de R) de la primera a la segunda muestra se conoce por el curioso nombre de **encogimiento de la validez**. El problema del encogimiento de la validez –y, por lo tanto, la necesidad de validez cruzada– puede ser de especial seriedad cuando la muestra inicial es pequeña. El problema disminuye conforme el tamaño de la muestra aumenta.

Esta operación del azar no es exclusiva de la correlación múltiple. Invitamos a regresar al capítulo 4, donde discutimos la confiabilidad de las diferencias: si examinamos un número suficiente de diferencias entre puntuaciones, casi es seguro que encontraremos algunas "diferencias significativas" sólo por azar. Volveremos a ver este fenómeno en el

capítulo 6 al elegir reactivos para una prueba basada en información del análisis de reactivos. El problema de sacar provecho del azar es pernicioso cuando tratamos con muchas variables.

Predicción estadística frente a predicción clínica

En la sección previa, describimos la metodología estadística para combinar información. Con las técnicas de correlación múltiple, determinamos de manera empírica qué información usar, qué pesos aplicar a lo que usamos y qué información descartar. Otra alternativa es combinar la información basada en la intuición y experiencia clínica. ¿Qué método es mejor para combinar información: el estadístico o el clínico? En algunas fuentes se denomina a esto la cuestión de la estadística contra la clínica, mientras que en otras se llama la cuestión de lo actuarial contra lo clínico. Consultar a los expertos clínicos no está limitado a los psicólogos clínicos, sino que incluye cualquier tipo de profesión, por ejemplo, consejeros o expertos en justicia criminal. Consideremos los siguientes dos escenarios.

Primero, queremos predecir el GPA de un grupo de 100 estudiantes de primer año. Podemos hacer una predicción estadística basada en los rangos de bachillerato y las puntuaciones del SAT utilizando la metodología de la correlación múltiple. También podemos pedirle a un grupo de consejeros de admisión que haga predicciones. Los consejeros tienen el rango de bachillerato e información del SAT; también tienen los fólderes de los estudiantes con cartas de recomendación, transcripciones de los cursos de bachillerato y registros de actividades y trabajo extracurricular. Los consejeros pueden combinar toda esta información de la manera que deseen y hacer un juicio clínico sobre el probable éxito, definido como el GPA. ¿Qué predicción será más exacta: la puramente estadística, basada en la regresión múltiple, o la clínica, basada en el uso intuitivo de la información?

Aquí está el segundo escenario. Tenemos un grupo de 50 pacientes en un hospital estatal; la mitad de ellos ha sido diagnosticada con el padecimiento A y la otra mitad, con el padecimiento B. Estos diagnósticos se basan en una evaluación extensa y en entrevistas múltiples con varios psicólogos. Tenemos mucha confianza en que los diagnósticos son correctos, pero ahora queremos ver cómo podemos clasificar con exactitud a estos individuos por medio de a) métodos estadísticos y b) entrevistas clínicas. Desarrollamos una ecuación de regresión múltiple sólo con el perfil de puntuaciones del MMPI para obtener la predicción estadística de pertenencia grupal, A frente a B. Tenemos tres psicólogos clínicos que entrevistan a los pacientes para tomar una determinación. Además de la información de las entrevistas, los clínicos también tienen las puntuaciones del MMPI. ¿Cuál será mejor predicción de la pertenencia grupal: la fórmula estadística o el juicio clínico?

Se han realizado numerosos estudios con un diseño similar al de estos dos escenarios. En general, las predicciones estadísticas son iguales o, muchas veces, mejores que las predicciones clínicas. Los clínicos hacen muecas frente a estos hallazgos, y los

estadísticos sienten vértigo. Meehl (1954) fue el primero en documentar la superioridad de las predicciones estadísticas sobre las clínicas en varios estudios. Otros informes han confirmado suficientemente este resultado. Dawes (1994, en especial el capítulo 3) presenta una revisión completa y amena de la literatura sobre este tema. ¿Podemos sustituir a los clínicos con fórmulas? A veces sí, a veces no. El desarrollo de fórmulas requiere de una base de datos adecuada en la que podamos apoyarnos, pero no siempre la tenemos. En ese caso, debemos apoyarnos en el juicio clínico para hacer lo mejor posible en cada situación. Además, necesitamos clínicos para desarrollar nociones originales de lo que se debe medir para dedicarnos a las fórmulas. También puede haber situaciones en que el juicio clínico, guiado *con firmeza* por las fórmulas estadísticas, puede ser mejor que las fórmulas por sí mismas. Grove y Meehl (1996), Grove, Zald, Lebow, Snitz y Nelson (2000), Kleinmuntz (1990) y Swets, Dawes y Monahan (2000) presentan una discusión detallada de este tema.

Teoría de la decisión: conceptos y términos básicos

La teoría de la decisión es un cuerpo de conceptos, términos y procedimientos para analizar los efectos cuantitativos de nuestras decisiones. Aplicada a la evaluación, la decisión implica usar pruebas, sobre todo en el contexto de la validez de criterio, para propósitos como selección, certificación y diagnóstico. Al aplicar la teoría, por lo general queremos optimizar los resultados de nuestras decisiones de acuerdo con ciertos criterios, los cuales pueden implicar ahorros en costos o tiempo. Las aplicaciones formales de la teoría de la decisión se vuelven con rapidez bastante complejas en términos matemáticos y están más allá del alcance de este libro. Sin embargo, una sencilla introducción de algunos conceptos y términos básicos de esta teoría ayudará en nuestra reflexión sobre la validez, en especial la validez de criterio.

Aciertos, positivos falsos y negativos falsos

Primero, vamos a familiarizarnos con las nociones de aciertos, positivos falsos y negativos falsos. Observemos el orden de los datos de la figura 5-7. Como en la figura 5-3, empleamos una prueba de admisión a la universidad para predecir el GPA de estudiantes de primer año. En muchas universidades, tener un GPA debajo de 2.0 resulta en una situación académica condicional; así que 2.0 es un punto de corte natural para el GPA, que en este caso es el criterio o variable *Y*. La prueba de admisión es la variable *X*, y tiene una media de 50 y una desviación estándar de 10. Decidimos usar una puntuación de 40 para elegir estudiantes para la clase de recién ingresados del próximo año. Así, 40 es el punto de corte.

Un **acierto** es un caso que tiene el mismo estatus respecto de la prueba y el criterio; es decir, los aciertos incluyen casos que excedieron el punto de corte del criterio y de la prueba (acierto positivo), así como casos que estuvieron por debajo del punto de corte del criterio y de la prueba (acierto negativo). Estos casos se ubican en los cuadrantes superior derecho e inferior izquierdo de la figura 5-7. Evidentemente, un índice alto de aciertos indica una buena validez de criterio de la prueba; sin embargo, a menos que la correlación entre prueba y criterio sea perfecta (1.00), habrá algunos errores en las predicciones. Los errores se clasifican de la siguiente manera. Los **positivos falsos** son casos que superan el punto de corte pero no se ajustan al criterio; estos casos se encuentran en el cuadrante inferior derecho de la figura 5-7. Los **negativos falsos** son aquellos en que la puntuación de la prueba está debajo del punto de corte, pero tienen éxito en el criterio; estos casos se encuentran en el cuadrante superior izquierdo de la figura 5-7.

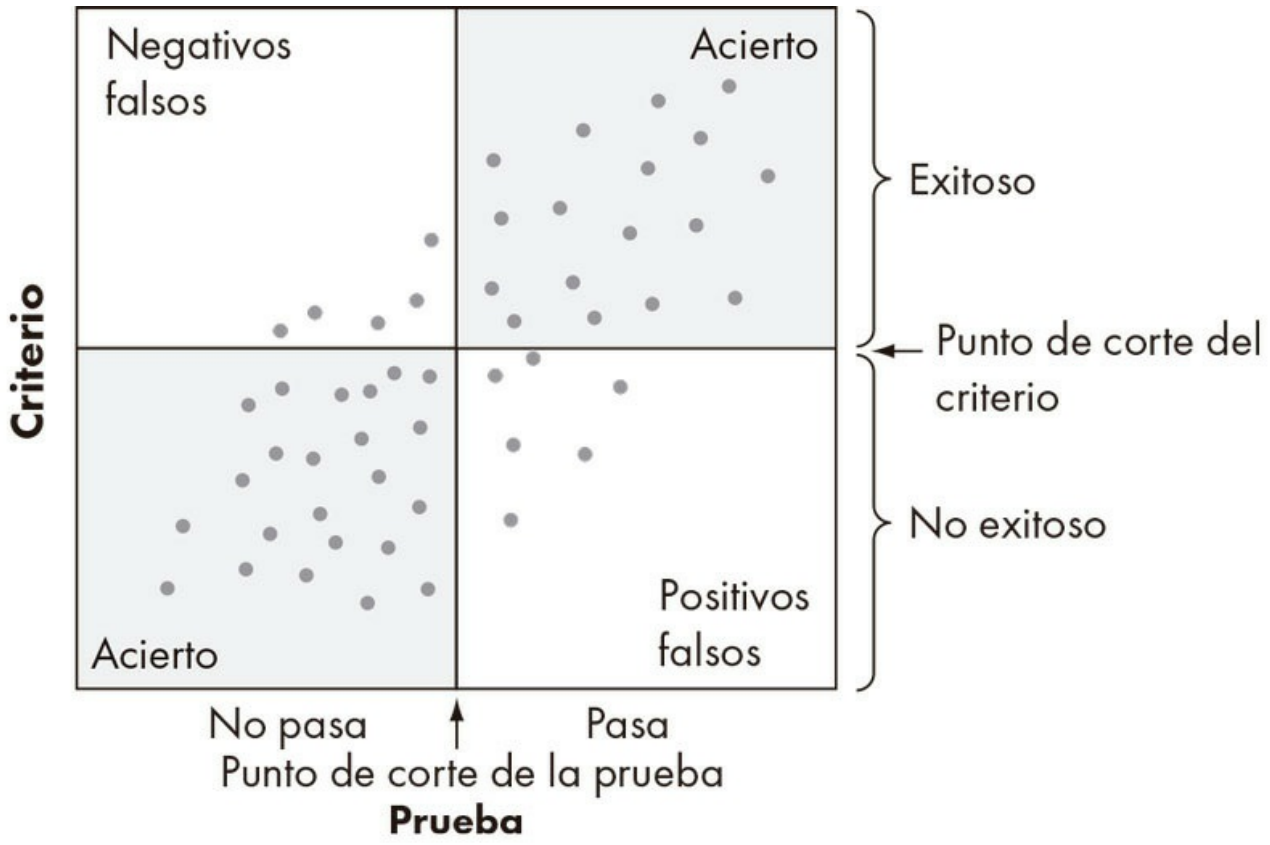


Figura 5-7. Aciertos, positivos falsos y negativos falsos en la relación entre prueba y criterio.

Nota: Es fácil que los psicómetras novatos ubiquen los "aciertos" en una gráfica como la de la figura 5-7, pero confunden a menudo los positivos falsos y los negativos falsos. Aquí presentamos una manera de tener claros estos términos. Siempre dibuja la gráfica de modo que la prueba quede en el eje horizontal y el criterio en el eje vertical. Entonces dibuja las líneas de los puntos de corte. Ubica las zonas de "aciertos", lo cual es fácil. Para ubicar los positivos falsos y los negativos falsos, recuerda que en una línea numérica, los valores "positivos" siempre van a la *derecha* y los "negativos", a la *izquierda*. Así, en los dos cuadrantes que quedan, los positivos falsos van a la derecha y los negativos falsos a la izquierda. Desafortunadamente, no todas las fuentes ubican la prueba en el eje horizontal y el criterio en el vertical, sino que las invierten, lo cual provoca cambios en el orden de los cuadrantes y migrañas en las personas que intentan comparar gráficas de distintas fuentes.

Dos factores afectan los porcentajes de aciertos, positivos falsos y negativos falsos. El *primer* factor es el grado de correlación entre la prueba y el criterio. Los casos extremos son los de correlación perfecta o de correlación cero. En el caso de una correlación perfecta, no habrá positivos falsos ni negativos falsos, pues todos serán aciertos. En el de una correlación cero, la suma de positivos falsos y negativos falsos será igual al número de aciertos.

El *segundo* factor es la ubicación de la puntuación de corte en la prueba. Los cambios en esta puntuación afectan el porcentaje relativo de positivos falsos y negativos falsos. Regresemos a la figura 5-7 para ver la ubicación de los positivos falsos y negativos falsos; ahora movamos un par de centímetros a la derecha el punto de corte. Los positivos falsos disminuirán, pero aumentarán los negativos falsos. Pero si desplazamos el punto de corte un par de centímetros a la izquierda de su posición original, veremos el efecto. La regla general aquí es que la correlación entre prueba y criterio no sea perfecta (en la práctica, siempre es así): hay una compensación entre el índice de de positivos falsos y el de negativos falsos. Al fijar el punto de corte de la prueba, el usuario de la prueba puede decidir qué resultado es preferible: un índice relativamente alto de positivos falsos o de negativos falsos. Por ejemplo, al usar una prueba para elegir buenos pilotos de líneas comerciales, podríamos estar interesados en minimizar los positivos falsos: los que pasan la prueba, pero no están calificados para volar. Esto resultaría en un número mayor de positivos falsos, es decir, los que están calificados para volar, pero no pasan la prueba. Por otro lado, en algunas circunstancias, podríamos querer minimizar los negativos falsos (p. ej., personas que probablemente cometan suicidio, pero con puntuaciones bajas en una prueba de tendencias suicidas), al mismo tiempo que se permite un aumento en los positivos falsos.

Índice base

El índice base es un concepto crucial para comprender la validez de una prueba, sobre todo en relación con el concepto de positivos falsos y negativos falsos. El índice base es el porcentaje de individuos de la población que tienen alguna característica.[2] Por ejemplo, el índice base de la esquizofrenia en la población general es de casi 1%, el de hombres solteros entre 25 y 29 años de edad es de 45% y el de adultos con grado de licenciatura es de casi 21%. Cuando el índice base es extremo, muy alto o muy bajo, es difícil mostrar que una prueba tiene una buena validez al identificar individuos del grupo meta. Consideremos una característica que sólo 0.5% de la población posee (1 de cada 200 personas). A menos que la prueba para identificar a a tales individuos tenga una validez excepcionalmente alta, minimizamos los errores en la clasificación simplemente declarando que nadie tiene la característica, sin importar la puntuación de la prueba. Una buena validez es lo más fácil de alcanzar cuando el índice base se acerca a 50%. Es importante notar que el índice base puede cambiar dependiendo de cómo se defina la población; por ejemplo, el índice base de un trastorno psicológico puede ser de 1% en la población general, pero de 30% en una población de personas que por voluntad propia busca ayuda en una clínica.

En una publicación clásica, Taylor y Russell (1939) explicaron cómo interactúa la validez de una prueba con los índices base de una razón de selección dada. Ofrecieron una descripción lúcida de la interacción, así como un conjunto de cuadros con valores selectos. Los cuadros de Taylor-Russell indican el grado de *mejora* en la selección que resulta del aumento en la validez de la prueba. En esta particular aplicación, necesitamos

un coeficiente de validez, una razón de selección conocida y un índice base. Por ejemplo, la razón de selección puede ser elegir 40% de candidatos a un trabajo o admitir 80% de aspirantes a la universidad. También necesitamos saber el índice base del éxito, por ejemplo, el porcentaje de casos que serían exitosos en un trabajo o en la universidad si no se utilizara ninguna prueba. Con esta información, los cuadros indican cuánto se puede mejorar usando una prueba con determinada validez en comparación con no usar ninguna prueba.

Sensibilidad y especificidad

Sensibilidad y especificidad son términos que tienen una estrecha relación con las nociones de positivos falsos y negativos falsos. Aplicamos estos términos cuando una prueba se usa para clasificar individuos en dos grupos, como alcohólicos y no alcohólicos o con riesgo suicida y sin riesgo suicida. Supongamos que queremos usar una prueba para identificar a personas con probabilidades de cometer suicidio; nuestro grupo criterio para validar la prueba es un grupo de personas que, en realidad, intentaron suicidarse. Tenemos un grupo de contraste de individuos que padecen depresión pero no han intentado suicidarse. Queremos una prueba y una puntuación de corte que a) identifique al grupo criterio (el de quienes han intentado suicidarse) y b) *no* identifique al grupo de contraste (el de quienes no han intentado suicidarse). La **sensibilidad** de una prueba es el grado en que identifica de manera correcta al grupo criterio, mientras que la **especificidad** es el grado en que la prueba *no* identifica o evita identificar al grupo de contraste. Sensibilidad y especificidad suelen expresarse como meros porcentajes. En el lenguaje de la sección previa, estos dos conceptos corresponden a los "aciertos".

Los datos del cuadro 5-11 ilustran la sensibilidad y especificidad de las distribuciones de las puntuaciones en una prueba de personas que han intentado suicidarse y de personas que no lo han intentado. Los cuatro ejemplos muestran dos grados diferentes de separación entre los grupos. Los ejemplos A y B muestran una buena separación, mientras que los ejemplos C y D muestran una separación menor. Los ejemplos también muestran dos diferentes puntos de corte en cada grado de separación y los cambios efectuados por las modificaciones en los puntos de corte. Al comparar los ejemplos A y B (en los cuales el grado de separación es el mismo), vemos que mover el punto de corte de +6 a +5 aumenta la sensibilidad de 74% a 88%, mientras que la especificidad disminuye. La *combinación* de sensibilidad y especificidad es mejor en los ejemplos A y B que en C y D, porque la separación de los grupos es más notable en A y B.

Cuadro 5-11. Distribuciones que ilustran los grados variables de sensibilidad y especificidad

	Ejemplo A		Ejemplo B		Ejemplo C		Ejemplo D	
Puntuación de la prueba	**Suicida**	**No suicida**	**Suicida**	**No suicida**	**Suicida**	**No suicida**	**Suicida**	**No suicida**
10	2		2		2		2	
9	4		4		4	1	4	1
8	5	1	5	1	5	3	5	3
7	11	3	11	3	11	2	11	2
6	15	2	15	2	15	8	15	8
5	7	8	7	8	7	14	7	14
4	3	14	3	14	3	11	3	11
3	2	11	2	11	2	4	2	4
2	1	4	1	4	1	5	1	5
1		5		5		2		2
0		2		2				
Sensibilidad	74%		88%		74%		88%	
Especificidad		88%		72%		72%		54%

Sensibilidad: 2, 4, 5, 11, 15 — Especificidad: 7, 3, 2, 1 (Suicida); 8, 14, 11, 4, 5, 2 (No suicida)

Hay 50 casos en cada distribución.
—★ indica el punto de corte.

¡Inténtalo!

En el ejemplo C, ¿cuál sería la sensibilidad y la especificidad si el punto de corte se fijara en 3+ en vez de 6+?

Los dos factores que afectan la sensibilidad y especificidad de una prueba son el grado de separación entre los grupos y la ubicación de los puntos de corte. Por lo general, mientras mayor es el grado de separación entre los grupos, mejor es la sensibilidad y la especificidad; es decir, mientras mejor discrimine la prueba entre los dos grupos, más alta es la sensibilidad y la especificidad. Con un grado fijo de separación entre los grupos, mover el punto de corte hará que sensibilidad y especificidad varíen de manera inversa, es decir, mientras una aumenta la otra disminuye.

Al considerar la discriminación entre grupos, es importante tener contrastes significativos; por ejemplo, es más útil contrastar los suicidas con los no suicidas que padecen depresión que contrastar suicidas con la población general. El primer contraste es más útil porque corresponde al tipo de distinción que solemos intentar hacer en la práctica; sin embargo, esta distinción puede llevar a una menor separación entre los grupos. Las distribuciones de las puntuaciones de los suicidas en comparación con las de los no suicidas que padecen depresión pueden ser muy parecidas a las del ejemplo C del cuadro 5-11, mientras que las distribuciones de los suicidas en comparación con la población general pueden ser más parecidas a las del ejemplo A. Sería posible obtener

una buena separación entre un grupo de suicidas y otro de individuos con una muy buena adaptación, pero ésa no es la clase de distinción que un clínico hace en su práctica cotidiana. Así, cuando se examinan los datos acerca de la sensibilidad, debemos estar atentos a la naturaleza de los grupos implicados en la comparación.

Como señalamos antes, sensibilidad y especificidad varían a la inversa con un grado fijo de separación. La pregunta natural es: ¿es mejor tener una sensibilidad relativamente alta y sacrificar un poco la especificidad? ¿O es mejor lo contario? Desde luego, es una pregunta similar a la de la compensación entre positivos falsos y negativos falsos; y la respuesta es la misma: depende. Depende de los riesgos y costos relativos y de otros factores implicados en la compensación. En el caso de los suicidas, probablemente optaríamos por aumentar la sensibilidad, pues preferimos identificar a la mayoría de suicidas –y ofrecerles ayuda– aunque eso signifique captar más no suicidas. En otras situaciones, podemos inclinarnos por lo contrario, es decir, disminuir la sensibilidad y aumentar la especificidad.

Nota: Para el psicómetra novato, los términos *sensibilidad* y *especificidad* son desafortunados, pues se ven y suenan de modo muy parecido aunque su significado casi sea opuesto. Para hacer más difícil la retención de los términos con claridad, algunas fuentes utilizan el término *selectividad* como equivalente de sensibilidad.

Las aplicaciones clínicas de las pruebas diagnósticas cada vez emplean más los conceptos de *poder predictivo positivo* (PPP) y *poder predictivo negativo* (PPN). Los cálculos de PPP y PPN surgen del mismo cuadro de cuatro cuadrantes de la figura 5-7 como selectividad y sensibilidad, pero configuran los datos de modo diferente, como se muestra a continuación:

- PPP = Verdaderos positivos/Todos los positivos
  o Verdaderos positivos/(Verdaderos positivos + Positivos falsos).
- PPN = Verdaderos negativos/Todos los negativos
  o Verdaderos negativos /(Verdaderos negativos + Negativos falsos).

Validez de constructo

Entre las categorías tradicionales de la validez (véase cuadro 5-1), la validez de constructo, al principio, es la más difícil de comprender. La noción básica de este tipo de validez puede describirse de la siguiente manera. Una prueba intenta medir algún constructo que, a veces, no tiene puntos obvios de referencia, como un cuerpo de contenido definido con claridad o un criterio externo. No obstante, se pueden aducir varias clases de evidencia para apoyar la proposición de que la prueba mide el constructo. La **validez de constructo** abarca todos estos métodos;[3] de hecho, empezando por esta línea de razonamiento, podemos pensar la validez de constructo como un concepto que incluye la validez de contenido y la validez de criterio. La correspondencia del contenido que está implicada en la validez de contenido y la correlación entre prueba y criterio son sólo casos –relativamente claros– que demuestran el grado en que la prueba mide el constructo. Las ediciones recientes de *Standards* no incluyen la validez de constructo como una categoría importante, sino que presentan diferentes métodos para demostrar la validez además de la validez de contenido y la de criterio. El cuadro 5-1 presenta una lista de otras fuentes de evidencia; en realidad, la lista de "otras" fuentes es interminable. Cualquier evidencia que de modo plausible apoye la proposición de que la prueba mide el constructo meta es pertinente. Sin embargo, hay ciertos tipos de evidencia que se vuelven a presentar en las discusiones sobre la validez de constructo, y nosotros presentamos esos tipos en esta sección.

Estructura interna

En el contexto de la validez, la consistencia interna significa lo mismo que cuando la tratamos al hablar de confiabilidad en el capítulo anterior (cf. pp. 89-93a»). Un nivel alto de consistencia interna, por ejemplo, una KR-20 o un coeficiente alpha altos, indica que la prueba mide *algo* de una manera consistente. Así, la consistencia interna alta apoya la afirmación de que la prueba mide un constructo o rasgo particular. A la inversa, es difícil mantener tal afirmación si la consistencia interna es baja.

La consistencia interna proporciona sólo una evidencia débil y ambigua en relación con la validez. Quizá lo mejor es pensar en la consistencia interna como un prerrequisito de la validez más que como evidencia de validez por sí misma. Una consistencia interna alta indica que un constructo se está midiendo, pero se requiere otra evidencia que sugiera de qué constructo podría tratarse.

Análisis factorial «127

El **análisis factorial** es una familia de técnicas estadísticas que ayudan a identificar las dimensiones comunes que subyacen en el desempeño en muchas medidas diferentes.

Estas técnicas se usan mucho en la construcción y validación de pruebas; tienen un papel destacado en los inventarios de personalidad y en las pruebas de inteligencia. De hecho, el desarrollo de la metodología del análisis factorial tiene una relación íntima con los debates clásicos sobre la naturaleza y medición de la inteligencia. Es difícil comprender el mundo de la evaluación de la personalidad o de la inteligencia sin conocer un poco del análisis factorial. Cuando decimos que se trata de una "familia" de técnicas, usamos esta palabra en un sentido amplio –como de la familia extensa– para incluir lo que se denomina análisis de los componentes principales, varios procedimientos de rotación, reglas de interrupción y temas relacionados.

Las técnicas del análisis factorial pueden volverse bastante complejas, por lo que una exploración detallada está más allá del alcance de este libro. Sin embargo, podemos resumir su propósito y el método general sin atorarnos en los detalles. Bryant y Yarnold (1995) presentan una excelente descripción semitécnica del análisis factorial, mientras que Tabachnick y Fidell (2007) ofrecen un detallado tratamiento técnico del tema.

El análisis factorial, como todas las técnicas estadísticas, empieza con los datos crudos; sin embargo, desde el punto de vista práctico, podemos pensar el análisis factorial empezando con una matriz de correlaciones. Consideremos las correlaciones del cuadro 5-12. Las variables *A* y *B* tienen una intercorrelación de .95, así que bien podríamos también hablar de una dimensión subyacente en estas dos variables. No es útil ni económico pensar en dos variables *diferentes* aquí. Ahora extendamos el caso a cuatro variables. La correlación r_{CD} también es muy alta, .93; otra vez, consideremos que es una sola variable o dimensión. Las *r'* de *A* y *B* con *C* y *D* son bastante bajas, por ejemplo, $r_{AC} = .20$, así que no podemos conjuntar *A* y *B* con *C* y *D*. Pero volvimos a empezar con cuatro variables y concluimos que hay, en realidad, sólo dos dimensiones subyacentes. Esto es, a nivel intuitivo, lo que hace el análisis factorial. Si expandimos este caso a 20 variables y todas sus interrelaciones, podemos ver que nuestra capacidad para seguir la pista de las cosas muy pronto se deterioraría. En estos casos más extensos, necesitamos la ayuda de procedimientos matemáticos formales, es decir, los del análisis factorial.

Cuadro 5-12. Matriz de correlaciones muestra de la discusión sobre el análisis factorial

Variable	***A***	***B***	***C***	***D***
A	—	.95	.13	.03
B		—	.20	.17
C			—	.93
D				—

Es útil construir algunas versiones geométricas de lo que hace el análisis factorial. Consideremos los ejemplos de la figura 5-8. En la distribución bivariada de la izquierda, las puntuaciones de las pruebas A y B tienen una correlación tan alta que necesitamos sólo una dimensión –el vector *AB*– para describir el desempeño. Es decir, no necesitamos

dimensiones *A* y *B* separadas. En el ejemplo de la derecha, *sí* necesitamos dos dimensiones –los vectores *A* y *C*– para describir la disposición de las puntuaciones. Aquí empleamos los términos *dimensión* y vector como equivalentes del término *factor* del "análisis factorial". Esta representación geométrica muestra, más o menos, cómo operan los procedimientos matemáticos del análisis factorial.

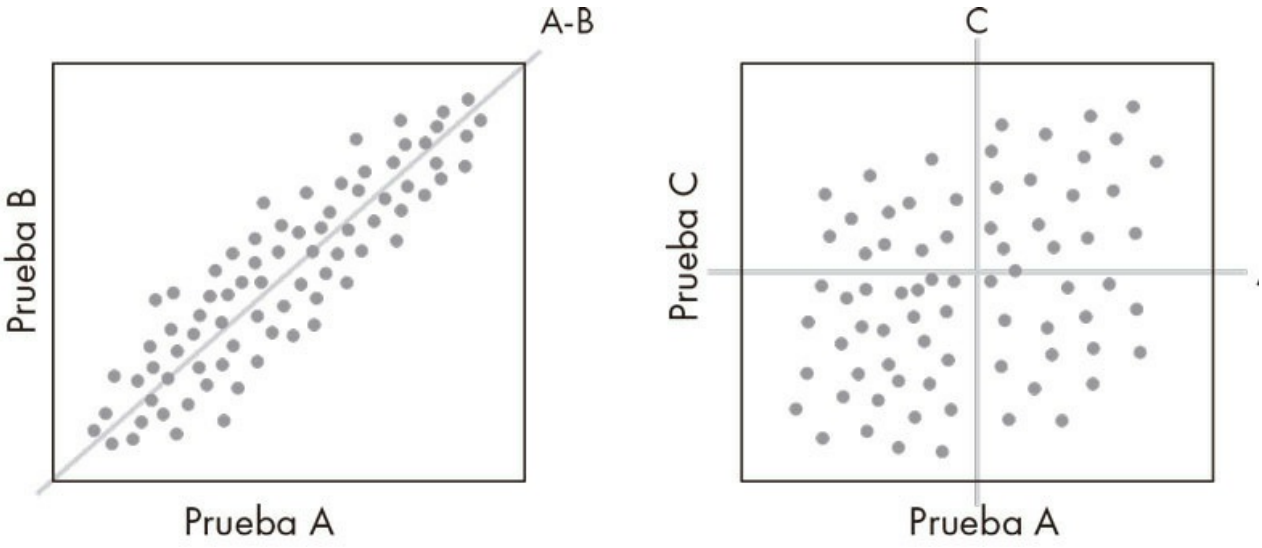


Figura 5-8. Análisis factorial representado en forma geométrica.

Los resultados del análisis factorial suelen representarse como una matriz factorial, que muestra el peso que cada variable original tiene en los factores recién establecidos. Los pesos son, en realidad, correlaciones entre las variables originales y los factores. El cuadro 5-13 presenta una matriz factorial muestra. En el lenguaje del análisis factorial, estas correlaciones se denominan "cargas". Lo habitual es considerar las cargas que superan el .30 como notables; desde luego, mientras mayores sean las cargas, más notables son.

Cuadro 5-13. Matriz factorial muestra

Prueba	**Factor I**	**Factor II**	**Factor III**
Vocabulario	.78	.07	.22
Lectura	.83	.13	.06
Problemas aritméticos	.54	.36	.16
Cálculo	.23	.89	.21
Estimación cuantitativa	.46	.67	.09

Los factores se "nombran" e interpretan de manera racional; por ejemplo, al examinarlo, el cuadro 5-13 sugiere que el factor I es una dimensión verbal, porque sus cargas más grandes están en Vocabulario y Lectura. Por su parte, el factor II parece ser

una dimensión cuantitativa. Es revelador que Problemas de aritmética cargue mucho más en el factor I que en el factor II, pues sugiere que comprender la presentación verbal de los problemas tiene mayor influencia para determinar la puntuación de una persona que la habilidad en matemáticas.

El factor III puede descartarse porque no es significativo: nada tiene una carga alta en él. Así, parece que dos dimensiones subyacentes explican este conjunto de cinco pruebas.

Este último ejemplo ilustra el uso del análisis factorial con pruebas enteras, pero también se aplica a los reactivos. Los resultados muestran cómo se agrupan los reactivos por sí mismos de acuerdo con las dimensiones subyacentes. Esta técnica se usa mucho en la construcción e interpretación de inventarios de personalidad e intereses. En el capítulo 6 se puede encontrar la discusión sobre este punto.

Hay muchas maneras diferentes de "extraer" los factores que difieren en los criterios matemáticos que emplean. Después de que se extraen los factores, lo habitual es "rotar" los ejes con el fin de facilitar la interpretación, para lo cual también disponemos de varios procedimientos. La rotación varimax es la más común. También hay diversos criterios para decidir cuándo dejar de extraer factores; los procedimientos computacionales del análisis factorial y la rotación son complicados. Sin embargo, es fácil realizarlos con un paquete estadístico como SPSS o SAS. No es fácil interpretar los resultados, pero sí lo es a menudo –de hecho es divertido– descifrar qué factores hay. En este capítulo, no estamos interesados en todos los procedimientos posibles para llevar a cabo el análisis factorial. El punto importante es que esta familia de técnicas nos ayuda a entender la estructura de las pruebas. De ahí que las técnicas son una fuente importante de información para la validez de constructo. Por ejemplo, los resultados del análisis factorial pueden sugerir que una medida de depresión en realidad tiene dos factores: uno definido por reactivos relacionados con el componente emocional y otro, primordialmente, por reactivos relacionados con los indicadores conductuales.

Procesos de respuesta

El estudio de la manera en que los examinados emprenden la tarea de responder una prueba, sus **procesos de respuesta**, puede proporcionar evidencia relacionada con la validez de la prueba. Por ejemplo, al estudiar una prueba de razonamiento cuantitativo, puede ser útil saber que el examinado suele pasar por diversas etapas para llegar a la respuesta en vez de aplicar una fórmula memorizada. Podríamos determinar que el examinado empleó un método de múltiples pasos si usamos una aplicación "en voz alta" de la prueba. Para investigar una prueba que pretende medir la capacidad de pensamiento creativo, una aplicación en voz alta puede ayudar a apoyar el argumento de que la prueba mide flexibilidad de pensamiento más que la mera riqueza de vocabulario.

El estudio de los procesos de respuesta también puede valerse de registros mecánicos o electrónicos. Por ejemplo, Exner (2003) informó su estudio de los movimientos oculares mientras los examinados respondían la prueba de manchas de Rorschach. Los resultados

aportaron conocimientos sobre la manera en que los examinados se aproximan a estímulos ambiguos como las manchas de tinta.

La evidencia proveniente de los procesos de respuesta no suele ofrecer evidencia sólida ni muy persuasiva en relación con la validez de la prueba. Además, dicha evidencia no se usa mucho para establecer la validez. Sin embargo, estudiar los procesos de respuesta a veces proporciona ideas útiles acerca de lo que la prueba puede o no estar midiendo.

Efecto de las variables experimentales

El efecto de las variables experimentales puede ayudar a demostrar la validez de una prueba. Consideremos estos ejemplos. Queremos establecer la validez del *Scranton Test of Anxiety* (STA); para ello, aplicamos la prueba a un grupo de 25 individuos, los sometemos a una situación que genere ansiedad y, luego, volvemos a aplicar el STA. Esperaríamos que las puntuaciones aumenten (pues indicarían ansiedad). Queremos establecer la validez del *Bechtoldt Creativity Test* (BCT [Prueba Bechtoldt de Creatividad]); para ello, aplicamos el BCT a 50 individuos, les damos 10 horas de instrucción en técnicas de pensamiento creativo y, luego, volvemos a aplicar el BCT. Esperaríamos que las puntuaciones aumenten. En ambos estudios, debemos tener grupos control para descartar la posibilidad de que cualquier aumento en las puntuaciones se deba a los efectos de la práctica.

Estudiar los efectos de variables experimentales es similar al método de grupos contrastados que tratamos antes al hablar de la validez de criterio. De hecho, desde el punto de vista lógico son lo mismo. Los estudios de grupos contrastados suelen emplear grupos que se forman por sí mismos (p. ej., personas deprimidas y no deprimidas), mientras que los grupos que tratamos al hablar de la validez de constructo se crean específicamente para estudiar la validez.

Cambios maduracionales o en el desarrollo

Otra potencial fuente de información con respecto de la validez de constructo son los cambios maduracionales o en el desarrollo. Esperamos que los niños, a lo largo de sus etapas, tengan una capacidad mental mayor. Mostrar que una capacidad mental refleja esta evolución ayuda a establecer la validez de la prueba. Sin duda, quedaríamos perplejos si una prueba de capacidad mental muestra las mismas puntuaciones en promedio para niños de 8, 9 y 10 años de edad. Uno de los principales métodos que usó Binet para defender la validez de los reactivos de su prueba fue demostrar cambios en las puntuaciones promedio de niños de distintas edades.

El aumento en las puntuaciones de las pruebas y en el desempeño en reactivos individuales en orden creciente de dificultad sirvió para defender la validez de las pruebas de aprovechamiento. Esperamos que el desempeño en lectura o matemáticas aumente del tercer grado al cuarto y quinto, y así sucesivamente. El estudio de los cambios en el

desarrollo, como el del efecto de las variables experimentales, puede pensarse como una variante del método de grupos contrastados. En este caso, contrastamos grupos de diferentes edades o grados.

Ya hemos revisado distintos procedimientos que ayudan a establecer la validez de constructo de una prueba. Como señalamos antes, la lista de maneras posibles para hacerlo es interminable. Cualquier evidencia que nos convenza de que la prueba mide el constructo meta es pertinente y útil.

Resumen de puntos clave 5-2

Algunas maneras importantes de estudiar la validez de constructo
Estructura interna
Análisis factorial
Procesos de respuesta
Efecto de las variables experimentales
Cambios maduracionales o en el desarrollo

Validez consecuencial

La **validez consecuencial** relaciona la prueba con las consecuencias de su uso e interpretación. El concepto incluye las consecuencias que se buscaban y las que no. ¿Cuáles son las consecuencias, resultados o implicaciones de usar una prueba? Por ejemplo, ¿cuáles son las consecuencias del uso sistemático de una prueba de admisión a la universidad? ¿Cuáles serán los resultados inesperados? Podemos notar que estas preguntas son diferentes de las que interrogan la utilidad de la prueba para predecir el GPA de los estudiantes de primer año. Podríamos preguntar si la prueba mejora (o quita mérito a) la calidad de la enseñanza en la universidad donde se aplica; también podríamos preguntar cuál es el efecto de pedir la prueba a los estudiantes de bachillerato que deben hacerla. Aquí hay otro ejemplo; supongamos que usamos una prueba para identificar a estudiantes para los cursos especiales de matemáticas. Podríamos preguntar si la prueba cubre el contenido del programa de matemáticas. Ése es un asunto de validez de contenido. También podríamos preguntar si el uso de la prueba trae beneficios educativos para los estudiantes identificados para tomar los cursos especiales. Ésa sería una pregunta acerca de la validez consecuencial.

Al menos dos temas separados necesitan considerarse aquí. El primero se relaciona con las declaraciones explícitas de algunos autores de pruebas respecto de las consecuencias. El segundo tema se relaciona con las posibles consecuencias independientemente de estas declaraciones. Consideremos el primer tema; recordemos que la validez de una prueba se define con respecto a su propósito, el cual es formulado por los autores. Si el propósito explícito incluye una consecuencia buscada, sin duda el proceso de validación debe tomar en cuenta esa consecuencia. Por ejemplo, si los autores de la prueba de admisión a la universidad afirman que el uso de la prueba mejorará la enseñanza en la institución que la aplique o hará que los estudiantes sean más diligentes, debe reunirse evidencia de validez relacionada con estas afirmaciones. Supongamos que los autores de un inventario de depresión aseguran que no sólo es una medida válida de depresión, sino que también trae una terapia más eficaz. Entonces, debe proporcionarse evidencia relacionada con una mejor terapia producto del uso del inventario.

El segundo tema es más complicado. Supongamos que los autores no hacen ninguna declaración respecto de las consecuencias, por ejemplo, acerca de mejorar la enseñanza o de influir en los estudiantes. Los autores sólo afirman que la prueba es útil para predecir el GPA. ¿Las mejoras en la enseñanza y la diligencia de los estudiantes, como consecuencias del uso de la prueba, aún son un asunto de validez? Además, supongamos que los autores de la prueba sólo hablan de la predicción del GPA, pero el director de la universidad afirma que el uso de la prueba, al final, mejora la calidad de la enseñanza en la universidad.

¿Esta afirmación del director convierte el asunto de las consecuencias en un tema de validez? Si es así, ¿quién es el responsable de reunir evidencia de la validez?

> ... es importante distinguir entre evidencia que es pertinente a la validez y la que puede servir para tomar decisiones informadas acerca de las políticas sociales, pero que fuera del reino de la validez.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Tanto el término como el concepto general de validez consecuencial son relativamente recientes en el vocabulario psicométrico. El término no apareció en la edición de 1985 de *Standards for Educational and Psychological Tests*. La edición de 1999 introdujo el término y, de hecho, le dedicó una sección completa a este concepto; el nuevo *Standards* extendió el espacio destinado para este término. Messick (1993) hizo el primer desarrollo sistemático de la noción de validez; él argüía que es una evolución importante en nuestra concepción de la validez de las pruebas. De manera incidental, también proporcionó un resumen conciso y útil de la evolución general de nuestras ideas acerca de todos los tipos de validez que vimos en este capítulo.

De ningún modo hay consenso con respecto al lugar que ocupa la validez consecuencial en el terreno de la psicometría. Algunas autoridades concuerdan con Messick en que es esencial (p. ej., Linn, 1997; Shepard, 1997), pero otras sienten que las consecuencias son un asunto de política y de políticas públicas, no de validez (p. ej., Mehrens, 1997; Popham, 1997); otros más sopesan si la validez consecuencial es un concepto psicométrico legítimo y cómo se reuniría evidencia pertinente (p. ej., Green, 1998; Lane, Parke, & Stone, 1998; Linn, 1998; Moss, 1998; Reckase, 1998; Taleporos, 1998). Por ejemplo, en el caso de la prueba de admisión a la universidad, ¿podemos identificar razonablemente todas las consecuencias de usarla? Y de las consecuencias que se pueden identificar, ¿cómo juzgamos si la suma de todas las posibles consecuencias es saludable o perniciosa? Lo que sí parece claro es que el debate sobre la validez consecuencial continuará por algún tiempo.

Apartados por completo del debate en los círculos académicos, Cizek y colegas (Cizek, Bowen, & Church, 2010; Cizek, Rosenberg, & Koons, 2008) han mostrado que la validez consecuencial es, en esencia, ignorada por las editoriales y los autores. Es decir, desde el punto de vista práctico, nadie puede imaginar qué hacer al respecto. Sin embargo, en el capítulo 16, veremos cómo algunos casos de tribunal han empleado las consecuencias para tomar decisiones legales.

Sesgos de las pruebas como parte de la validez

El sesgo de las pruebas (o lo opuesto, su neutralidad) se refiere a si la prueba mide el constructo meta de manera equivalente en diferentes grupos. Una prueba sesgada no lo hace así, pero una prueba neutral sí. Ya que esta pregunta se ocupa de cómo se mide el constructo, por lógica entra en este capítulo sobre validez. Sin embargo, hemos pospuesto tratar este tema hasta el final del capítulo 6 sobre la elaboración de pruebas, porque durante este proceso se realizan los esfuerzos para tratar con los sesgos. Desde el punto de vista práctico, necesitamos saber acerca de la elaboración de pruebas para comprender algunos métodos clave para examinar el sesgo. Así, al menos en este caso, dejaremos que las consideraciones prácticas prevalezcan sobre las lógicas.

Preocupaciones prácticas

Para este momento, debería estar claro que la validez de las pruebas no es un asunto sencillo. Hay numerosas maneras de estudiarla y cada una tiene sus limitaciones. En el caso de muchas pruebas, se han llevado a cabo un gran número de estudios con resultados variables. Puede haber razones legítimas para que la validez de una prueba varíe de una situación a otra; por ejemplo, la depresión puede tener un aspecto diferente en adultos jóvenes y en adultos mayores, lo cual altera la validez del *Scranton Depression Inventory* para estos grupos. El *Western Admissions Test* puede tener una validez algo diferente en el Colegio Ivy y en la Universidad Estatal de Behemoth, pero no debido a las diferencias en la heterogeneidad grupal, sino a las diferencias en los cursos de las dos instituciones. En esta sección final del capítulo, trataremos de formular algunos consejos para tratar con estos temas.

Integración de la evidencia

En el análisis final, el usuario profesional de pruebas debe sopesar toda la evidencia disponible y hacer juicios acerca de la probable validez de una prueba utilizada en ciertas circunstancias. El proceso de sopesar toda la evidencia y juzgar la pertinencia de los estudios existentes para un uso específico anticipado se denomina **generalización de la validez**. La aplicación inteligente de la generalización de la validez requiere a) conocimiento del área de contenido pertinente (p. ej., depresión, aprovechamiento de la lectura, desempeño en la universidad), b) estar familiarizado con la investigación realizada con la prueba y pruebas similares, c) comprensión de los conceptos y procedimientos tratados en este capítulo y d) análisis perceptivo de las circunstancias locales para un uso anticipado de la prueba.

> Un tema importante en los escenarios educativos y laborales es el grado en que la evidencia de la validez basada en las relaciones entre prueba y criterio puede generalizarse a una nueva situación sin más estudios de validez en esa situación... Los resúmenes estadísticos de los estudios de validación pasados en situaciones similares pueden ser útiles al estimar las relaciones entre prueba y criterio en una nueva situación. Se habla de esta práctica como estudio de generalización de la validez.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Standards trata la generalización de la validez como un subtema de "Evidencia basada en las relaciones con otras variables". Por ejemplo, debemos juzgar la semejanza de una universidad con otras donde la relación (validez predictiva) entre una prueba de admisión y el GPA de un estudiante de primer año ya se ha determinado. Sin embargo, el concepto de generalización de la validez se aplica a todo tipo de determinaciones de validez. Por ejemplo, el manual de una prueba informa los resultados de un análisis factorial e indica que la prueba parece medir cuatro factores distintos. Ese estudio se llevó a cabo con un

grupo particular de examinados: con una cierta distribución de edades, desglose del género, estatus socioeconómico y así sucesivamente. El usuario profesional de la prueba debe juzgar el grado en que esos resultados se aplican a una situación local.

Muchas de las pruebas más usadas, como el SAT, Rorschach y MMPI-2, han sido objeto de miles de estudios, algunos de los cuales se concentran en la validez, otros en la confiabilidad y otros más en otras características de las pruebas. Puede ser una tarea abrumadora resumir todos los estudios pertinentes sobre un aspecto de una sola prueba, por ejemplo, la validez del Rorschach.

Sin embargo, realizar tales revisiones es parte de la tarea de integración de la evidencia de la validez de una prueba. El **metaanálisis** es una técnica para resumir la información real estadística de muchos estudios diferentes sobre un solo tema. El resultado del metaanálisis es un estadístico como el coeficiente de correlación o una medida del tamaño del efecto que representa una generalización de todos los estudios sobre el tema. En la actualidad, el metaanálisis es la técnica preferida para resumir información, como la de validez o confiabilidad de una prueba, proveniente de diversos estudios. Los siguientes son ejemplos de metaanálisis realizados con pruebas muy usadas: Finger y Ones (1999), sobre la versión para computadora del MMPI; Hiller, Rosenthal, Bornstein, Berry y Brunell-Neulieb (1999), sobre el Rorschach y el MMPI; Kuncel, Hezlett y Ones (2001) y Kuncel, Wee, Searfin y Hezlett (2010), sobre el *Graduate Record Exmination; Shafer* (2006), sobre cuatro medidas de depresión; y Parker, Hanson y Hunsley (1988), sobre el MMPI, el Rorschach y el WAIS. Cualquiera de estos dará al lector una buena idea de cómo ayuda el metaanálisis con la generalización de la validez.

En el análisis final: un estándar relativo

En la conclusión del capítulo 4, preguntamos qué tan alta debía ser la confiabilidad. Aunque no hay una sola y definitiva respuesta, identificamos algunas pautas para responder a la pregunta. Ahora, en la conclusión de este capítulo sobre validez, debemos hacer una pregunta similar: ¿qué tan alta debe ser la validez? Desafortunadamente, la respuesta a esta pregunta es, por fuerza, menos definitiva aun que en el caso de la confiabilidad. En el análisis final, la respuesta es muy relativa; necesitamos preguntar si una prueba es más o menos válida que otra. Ambas pueden tener una validez baja, pero elegiremos la que sea relativamente mejor. A veces la pregunta práctica es si usar una prueba o nada, en cuyo caso, podemos estar satisfechos si la prueba muestra algún grado de validez. La alternativa puede ser basar la decisión en ninguna información, lo cual es equivalente a lanzar al aire una moneda. En la práctica cotidiana, tenemos que elegir la mejor prueba que podamos obtener, mientras nos esforzamos por desarrollar mejores fuentes de información.

A este respecto, vale la pena señalar que la validez de las pruebas psicológicas se compara favorablemente con la validez de muchas pruebas médicas usadas comúnmente (Meyer *et al*., 2001). A menudo nos sentimos decepcionados por no obtener coeficientes de validez más altos en las pruebas psicológicas. Muchas de nuestras pruebas médicas

tampoco son tan buenas. Tenemos que usar lo mejor que esté disponible a pesar de que esté lejos de ser perfecto.

Resumen

1. La validez se refiere al grado en que la interpretación de la puntuación de una prueba es apropiada para un propósito específico. Es la característica más importante de una prueba.
2. Los conceptos de subrepresentación del constructo y varianza irrelevante para el constructo son útiles cuando se considera el grado de superposición entre la prueba y el constructo que se pretende medir.
3. La validez aparente se refiere a si una prueba tiene el aspecto de ser válida. No es una demostración empírica de validez. Es útil para que el público acepte la prueba.
4. La validez de contenido se ocupa de la correspondencia entre el contenido de la prueba y un cuerpo de conocimientos o habilidades bien definidos. Se usa con las pruebas de aprovechamiento y reclutamiento.
5. La validez de criterio expresa la relación entre las puntuaciones de la prueba y el estatus de algún otro criterio que refleje el constructo de interés. El estatus del criterio puede determinarse aproximadamente al mismo tiempo que se aplica la prueba (validez concurrente) o en un momento posterior (validez predictiva).
6. En la validez de criterio, éste puede ser un criterio externo y factible, un grupo de contraste u otra prueba.
7. Cuando la validez de criterio se expresa como correlación (r_{XY}) entre prueba y criterio, llamamos a esta correlación coeficiente de validez. Una vez establecida r_{XY}, podemos usar este coeficiente de validez para predecir el estatus de un criterio a partir de la puntuación de la prueba. Además, podemos determinar el error estándar de estimación y usarlo para determinar las probabilidades relacionadas con la exactitud de la estimación.
8. Todos los factores que afectan la interpretación del coeficiente de correlación, incluyendo la linealidad, homocedasticidad y heterogeneidad grupal, también afectan la interpretación de los coeficientes de validez.
9. Las confiabilidades de la prueba y del criterio afectan el coeficiente de validez. Algunas fórmulas sencillas permiten corregir este coeficiente por una confiabilidad limitada (atenuada).
10. El criterio de interés debe definirse en términos operacionales. Por lo general, hay varias definiciones alternativas operacionales que deben tomarse en cuenta.
11. La contaminación del criterio se refiere a una situación indeseable en la que las puntuaciones de la prueba influyen en el criterio, de modo que el coeficiente de validez se infla de manera injustificada.
12. Validez convergente y discriminante son conceptos útiles cuando se piensa en la validez de criterio. La validez convergente significa que la prueba tiene una correlación alta con otra prueba o fuente de información que mide el constructo meta de la prueba. La validez discriminante significa que la prueba tiene una baja correlación con otras pruebas o fuentes de información que son indicadores de un constructo diferente.

13. La correlación múltiple es una técnica estadística para combinar información de diversas pruebas (u otras fuentes de información) y predecir el estatus de un criterio. Para obtener los mejores predictores posibles, los procedimientos de correlación múltiple asignan pesos a las pruebas de acuerdo con sus contribuciones únicas a la predicción. Los procedimientos de correlación múltiple son de especial importancia para estudiar la validez incremental. Donde existan bases de datos adecuadas, las fórmulas estadísticas serán, por lo general, iguales o mejores que los juicios clínicos al combinar información para tomar decisiones.

14. La teoría de la decisión es un conjunto de conceptos y procedimientos para analizar los efectos cuantitativos de las decisiones. Aplicada a la validez de criterio, la teoría de la decisión incluye los conceptos aciertos, positivos falsos y negativos falsos. La sensibilidad y especificidad de la prueba son otros dos conceptos útiles. Los índices base y la ubicación de los puntos de corte en la prueba tienen consecuencias importantes en este marco.

15. La validez de constructo se definió, en su origen, como un conjunto misceláneo de técnicas consideradas pertinentes para la validez distinta de la de contenido y de criterio. La validez de constructo ahora se considera un concepto global que abarca todos los tipos de evidencia de la validez, como la estructura interna de la prueba, en especial como se revela en el análisis factorial, el estudio de los procesos de respuesta, el efecto de las variables experimentales y los cambios de desarrollo. Cualquier tipo de evidencia que nos convenza de que la prueba mide, al menos hasta cierto punto, el constructo meta para un propósito particular puede considerarse como validez de constructo.

16. La validez consecuencial relaciona la prueba con las consecuencias últimas de su uso e interpretación. Esta noción abarca las consecuencias buscadas y las imprevistas. La validez consecuencial es una recién llegada a la discusión sobre la validez y despierta opiniones opuestas acerca de cómo tratarla con exactitud.

17. El usuario de la prueba debe integrar la evidencia de la validez de diversas fuentes para llegar a un juicio acerca del grado en que la prueba cumple con su propósito. En el análisis final, cuando llegamos a ese juicio, el usuario trata de responder a la pregunta: ¿es mejor que use esta prueba como fuente de información o no?

Palabras clave

acierto
análisis de puesto
análisis factorial
análisis multirrasgo-multimétodo
atenuación
coeficiente de validez
constructo
contaminación del criterio
correlación múltiple
criterio externo
ecuación de regresión múltiple
encogimiento de la validez
error estándar de estimación
especificidad
generalización de la validez
índice base
metaanálisis
negativo falso
positivo falso
procesos de respuesta
sensibilidad
subrepresentación del constructo
tamaño del efecto
validez
validez aparente
validación cruzada
validez concurrente
validez consecuencial
validez convergente
validez de constructo
validez de contenido
validez de criterio
validez discriminante
validez incremental
validez instruccional
validez predictiva
varianza irrelevante para el constructo

Ejercicios

1. Para reafirmar tu comprensión de los conceptos de subrrepresentación del constructo y varianza irrelevante para el constructo, haz un dibujo como el de la figura 5-1 que represente estos casos:

a. Un constructo incluye 12 componentes y una prueba abarca seis de ellos. Además, las puntuaciones de la prueba están ligeramente influidas por factores diferentes de estos seis componentes.

b. Los reactivos de una prueba buscan medir pensamiento creativo, pero la prueba, casi en su totalidad, es sólo una medida de vocabulario. (Supón, para este ejercicio, que vocabulario y pensamiento creativo tienen una correlación baja.)

2. Identifica un puesto de trabajo con el que estés familiarizado y crea un cuadro de especificaciones de él. Con este cuadro, examina la validez de contenido de una prueba diseñada para reclutar a alguien para ese puesto.

3. El coeficiente de validez del Western Admissions Test (WAT; variable *X*) para predecir el GPA de un estudiante de primer año (variable *Y*) es $r_{XY} = .60$. Éstas son las medias y desviaciones estándar de *X* y *Y*.

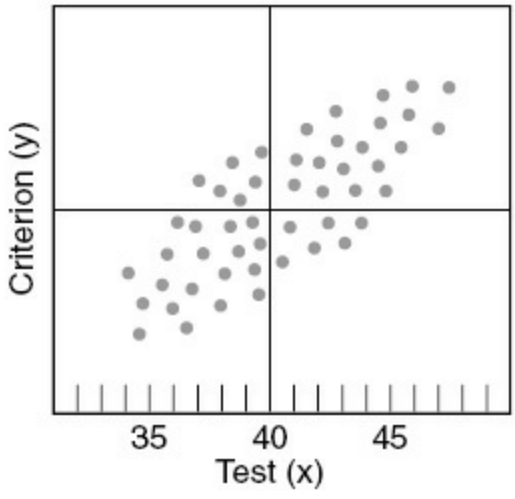
	M	*DE*
X	50	10
Y	3.00	.40

a. ¿Cuál es el GPA predicho a partir de una puntuación de 65 en WAT? Usa la fórmula 5-2.

b. ¿Cuál es el error estándar de estimación de estos datos? Usa la fórmula 5-3

c. ¿Cuál es la probabilidad de que una persona con una puntuación de 35 en WAT alcance un GPA por debajo de 2.00?

4. Hay 50 casos en esta distribución bivariada.



a. Con un punto de corte en *X* fijado en 40, como se muestra, ¿cuántos aciertos, positivos falsos y negativos falsos hay aquí?

b. Desplaza el punto de corte en *X* a un nivel más alto, digamos 42, de modo que el número de positivos falsos disminuya. Ahora cuenta el número de aciertos, positivos falsos y negativos falsos.

5. En los estudios sobre validez predictiva, el éxito en la universidad se define como el GPA al final del primer año. ¿Qué otras definiciones operacionales de "éxito en la universidad" podrías inventar? ¿Alguna de estas definiciones alternas llevaría a usar pruebas de admisión diferentes?

6. Las pruebas A y B pretenden predecir el GPA en la universidad. La prueba A es corta y tiene una confiabilidad de .60; su correlación con el GPA es .45. La prueba B es muy larga y tiene una confiabilidad de .95; su correlación con el GPA es .50. Aplica la corrección por falta de confiabilidad en las pruebas, no en el criterio, a las correlaciones de ambas pruebas con el GPA. ¿Cuáles son las correlaciones corregidas con el GPA de ambas pruebas? Con base en estos resultados, ¿concluirías que vale la pena revisar la prueba A para hacerla más confiable? ¿Cómo podrías hacerlo? (*Pista*: ve el capítulo 4 en la página 90b».)

7. Regresa a la figura 5-6. Por medio de un diagrama como los que hemos presentado aquí, ilustra esta descripción verbal. Las pruebas X, Y y Z pueden usarse para predecir el criterio C. Las pruebas X y Y tienen correlaciones altas entre sí y con C. La prueba Z tiene una correlación moderadamente baja con X y Y, así como con C.

8. Consulta fuentes electrónicas o impresas y revisa reseñas de cualquier prueba de una edición reciente del *Mental Measurements Yearbook* de Buros. ¿Qué dicen los autores de las reseñas acerca de la validez de la prueba? ¿Qué tipos de evidencia de la validez discuten?

9. ¿Qué usarías como *definición operacional de "éxito"* en cada una de estas ocupaciones?

Profesor universitario _____

Jugador de beisbol _____

Abogado _____

10. Con los datos del apéndice D1: GPA, usa el SPSS u otro paquete estadístico para determinar la correlación entre la puntuación total del SAT y el GPA. De acuerdo con la terminología de este capítulo, ¿cómo llamarías a este coeficiente de correlación? Después, crea una distribución bivariada (dispersograma) de los datos. Trata de usar la función ajuste de línea para generar la línea de regresión del dispersograma.

Notas

[1] En estos ejemplos, suponemos que los autores de informes interpretativos emplean evidencia empírica para hacer las afirmaciones. Sin embargo, algunos informes se basan en ilusiones o hipótesis más que en evidencias.

[2] La literatura clínica prefiere el término índice de prevalencia a índice base.

[3] Véase Kane (2001) para consultar un útil resumen de la evolución del concepto de validez de constructo.

CAPÍTULO 6
Elaboración de pruebas, análisis de reactivos y neutralidad

Objetivos

1. Enumerar los pasos para elaborar una prueba.

2. Identificar las cuestiones que se deben considerar en el diseño preliminar de una prueba.

3. Identificar ejemplos comunes de reactivos de respuesta cerrada.

4. Identificar ejemplos comunes de reactivos de respuesta abierta.

5. Citar algunos de los métodos para calificar reactivos de respuesta abierta.

6. Discutir los aspectos positivos de los reactivos de respuesta cerrada y abierta.

7. Dar ejemplos de algunas reglas para redactar reactivos de respuesta cerrada y, luego, de respuesta abierta.

8. Identificar los dos tipos principales de estadísticos tradiScionales de reactivos.

9. Describir los atributos de una curva característica de reactivo.

10. Citar las directrices para elegir reactivos.

11. Esbozar el conjunto de materiales que deben estar disponibles para la publicación de una prueba.

12. Definir qué significa neutralidad o sesgo de la prueba.

13. Identificar los tres métodos principales para investigar la neutralidad de la prueba.

Introducción «135-137a

En este capítulo se explica a grandes rasgos los pasos que, por lo común, se siguen para elaborar una prueba. El título del capítulo hace mención especial del "análisis de reactivos", porque estos procedimientos analíticos tienen un papel decisivo en la elaboración de pruebas; sin embargo, dichos procedimientos sólo son una parte de esta tarea. En este capítulo se describe cada uno de los seis pasos principales para elaborar una prueba,[1] los cuales se citan en la figura 6-1.

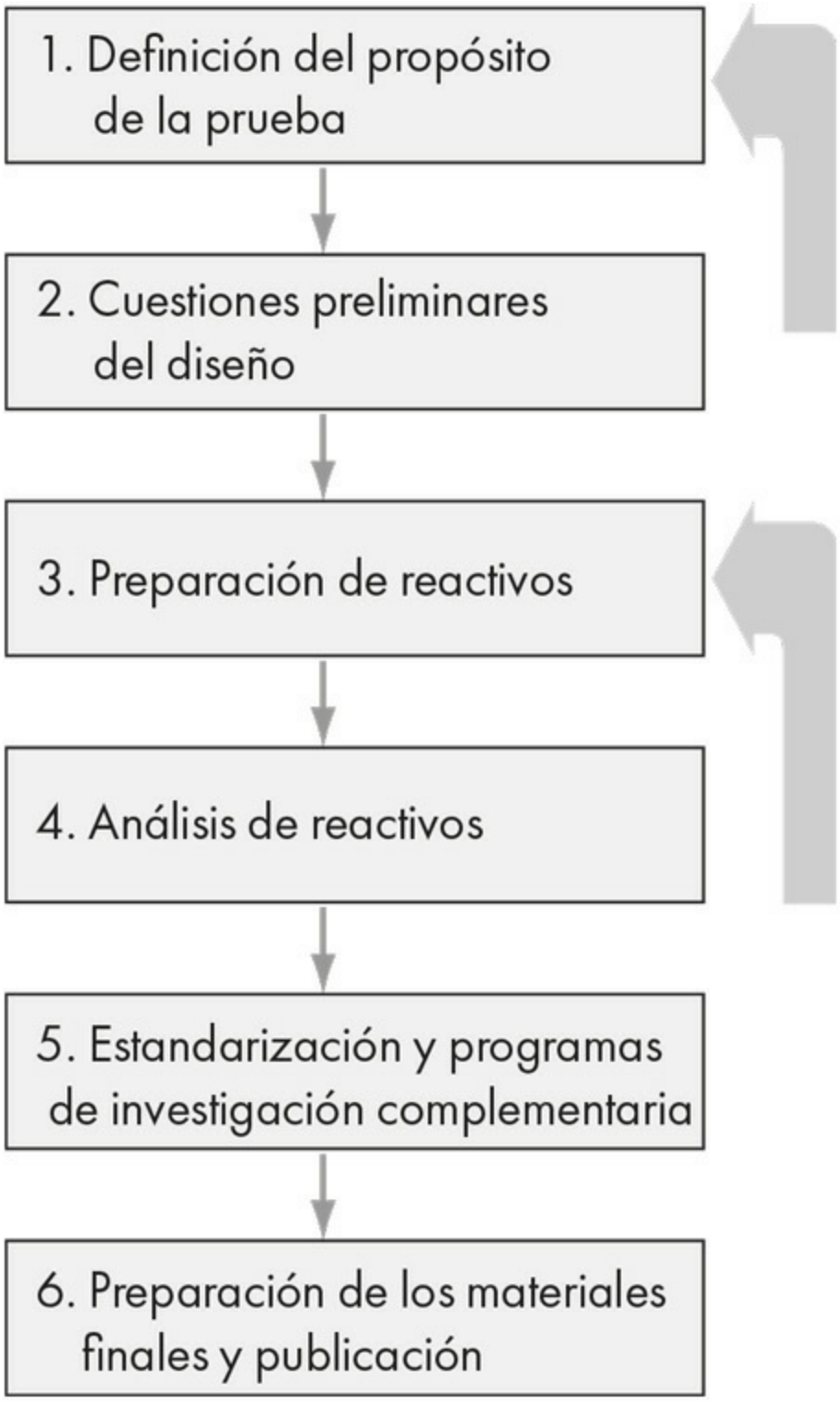


Figura 6-1. Principales pasos para elaborar una prueba.

Los pasos no siempre son, por completo, distintos; en la práctica, a menudo habrá

cierta superposición y reciclado entre ellos. Esto es particularmente cierto entre los pasos 1 y 2 y entre los pasos 3 y 4, como se verá con claridad en nuestra descripción. Sin embargo, esta lista de pasos nos da la progresión lógica y el orden cronológico típicos del trabajo de elaboración de una prueba.

> El proceso para elaborar pruebas educativas y psicológicas debe empezar citando el propósito o propósitos de la prueba, el usuario y los usos para los que está pensada, el constructo o dominio de contenido que se medirá y la población de examinados a la que está dirigida.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Después de describir la elaboración de pruebas y el análisis de reactivos, retomaremos la cuestión del sesgo: qué significa, cómo se estudia y qué procedimientos se emplean para tratar con él al aplicar pruebas.

Definición del propósito de la prueba

La elaboración de una prueba empieza con una clara **formulación del propósito** de la prueba, la cual incluye una descripción del rasgo que se medirá y del público al que está dirigida. El propósito debe formularse teniendo en mente la clase de interpretación que, en última instancia, se hará de las puntuaciones de la prueba. El cuadro 6-1 contiene formulaciones del propósito de varias pruebas muy utilizadas; los propósitos suelen formularse de manera bastante sencilla, a menudo con una sola oración.

Desde el punto de vista práctico, después de que se ha formulado con claridad el propósito de la prueba, *no* debemos continuar de inmediato con su construcción, pues el siguiente paso debe ser determinar si ya existe una prueba apropiada. Recurrir a las fuentes de información citadas en el capítulo 2 puede ayudar a tomar esta decisión. Construir una nueva prueba –al menos, una buena prueba– es una tarea larga, difícil y costosa. Una recomendación para las personas sensatas: tomen su tiempo para determinar si una prueba existente puede servir a sus propósitos antes de intentar crear una nueva.

Cuadro 6-1. Formulaciones del propósito de diversas pruebas muy utilizadas

"Los inventarios NEO son medidas concisas de las cinco dimensiones o dominios principales de la personalidad y los rasgos o facetas más importantes que definen cada dominio. En conjunto, las cinco escalas de dominios amplios y las 30 escalas de facetas específicas permiten realizar una evaluación integral de la personalidad adolescente y adulta." (McCrae & Costa, 2010, p. 1)
"La Escala Wechsler de Inteligencia para Adultos – IV (WAIS-IV) es un instrumento clínico integral de aplicación individual para evaluar la inteligencia de personas de 16 a 90 años de edad" (Wechsler, 2008a, p. 1)
"El Inventario Multifásico de Personalidad de Minnesota-2 (MMPI-2) es una prueba de amplio espectro diseñada para evaluar varios patrones importantes de trastornos de la personalidad y emocionales." (Hathaway & McKinley, 1989, p. 1)
"El Otis-Lennon School Ability Test, Octava Edición (OLSAT8)... está diseñado para medir las habilidades de razonamiento verbal, cuantitativo y figurativo que tienen las relaciones más estrechas con el aprovechamiento escolar." (Otis & Lennon, 2003, p. 5)

Resumen de puntos clave 6-1

Los primeros dos pasos decisivos en la elaboración de pruebas

Definir con claridad el propósito: variable(s) meta y grupo meta

Considerar cuestiones preliminares del diseño, como modo de aplicación, extensión, formato de los reactivos, entrenamiento, número de puntuaciones e informes de puntuaciones

¡Inténtalo!

Elige un rasgo de tu interés. Supón que vas a elaborar una prueba para medir ese rasgo. Formula el propósito de tu prueba incluyendo una definición de la población a la que está dirigida.

Cuestiones preliminares del diseño

En las primeras etapas de la elaboración de pruebas, el autor debe tomar varias decisiones acerca del diseño, las cuales se basan en el propósito de la prueba y la interpretación de las puntuaciones que se tienen pensadas, así como otras consideraciones prácticas. Deben considerarse las siguientes **cuestiones del diseño**:

- Modo de aplicación. ¿La prueba se aplicará de manera individual o también podrá aplicarse de manera grupal? La aplicación grupal será más eficiente, pero la individual permite más adaptabilidad del formato de los reactivos y la observación clínica del examinado.
- Extensión. ¿Aproximadamente cuánto tiempo se llevará la prueba? ¿Será corta con un tiempo aproximado de 15 minutos para su aplicación, o será más larga y tomará 45 minutos o varias horas? Las pruebas cortas son, obviamente, más eficientes, pero pueden significar una confiabilidad limitada y sólo una puntuación. La extensión no sólo es cuestión del número de reactivos y el tiempo de aplicación, sino que se relaciona íntimamente con el tema de qué tan sensible será la prueba. ¿La prueba será una medida general o global, del rasgo evaluado? ¿Proporcionará la base de un análisis diagnóstico sensible del rasgo?
- Formato de los reactivos. ¿Qué formato de reactivos se utilizará: opción múltiple, verdadero-falso, de acuerdo-en desacuerdo, respuesta abierta? Un formato de respuesta abierta permite obtener respuestas más ricas y de mayor flexibilidad, pero casi inevitablemente será más difícil de calificar y, por lo tanto, más costoso de usar.
- Número de puntuaciones. ¿Cuántas puntuaciones proporcionará la prueba? Esta cuestión, por necesidad, está relacionada con la de extensión de la prueba. Más puntuaciones permiten hacer interpretaciones adicionales, pero también exigen más reactivos y, por lo tanto, más tiempo de aplicación.
- Informes de las puntuaciones. ¿Qué clase de informes se producirán? ¿Habrá un registro simple hecho a mano de la puntuación o un conjunto elaborado de informes generados por computadora que puedan incluir informes interpretativos? ¿Exactamente qué será informado: sólo una puntuación total o también el desempeño en grupos de reactivos?
- Capacitación para la aplicación. ¿Cuánto entrenamiento se necesitará para aplicar y calificar la prueba? ¿Se necesitará una capacitación profesional amplia para aplicar, calificar e interpretar la prueba? Si así fuera, ¿cómo se ofrecería dicha capacitación?
- Investigación de los antecedentes. En la etapa del diseño preliminar, puede ser necesario llevar a cabo una investigación de los antecedentes del área correspondiente a lo que se pretende evaluar, a menos que se tenga un conocimiento detallado de ella. Esta investigación debe incluir una búsqueda de la

literatura pertinente. Si la prueba está pensada para tener una amplia aplicación práctica, la investigación debe incluir también discusiones con profesionales (p. ej., clínicos, consejeros, psicólogos escolares, etc.) de los campos en los que la prueba podría emplearse.

Muchos tratamientos de la elaboración de pruebas comienzan con la "redacción de reactivos", pero no se puede empezar con esto (al menos, no debe ser así) hasta que las cuestiones preliminares del diseño se hayan considerado minuciosamente, pues éstas determinarán qué clase de reactivos y cuántos de ellos tendrán que escribirse. Las malas decisiones acerca del diseño original de la prueba no podrán remediarse en las etapas de redacción o análisis de reactivos.

Las reflexiones relacionadas con las cuestiones preliminares del diseño pueden llevar a refinar la formulación del propósito de la prueba. Ésta es la razón de la flecha que lleva del paso 2 al paso 1 en la figura 6-1. Por ejemplo, la decisión de hacer una prueba más corta en vez de una más larga puede llevar a formular un propósito más circunscrito, o la discusión con los usuarios puede llevar a ampliar el público al que está dirigida la prueba.

¡Inténtalo!

Tomando como referencia la prueba cuyo propósito formulaste en el ejercicio anterior, en esta misma página, responde las siguientes preguntas acerca del diseño de tu prueba:

¿Cuántos reactivos tendrá la prueba?
¿Cuántas puntuaciones resultarán de ella?
¿Su aplicación será individual o grupal?
¿Cuántos minutos aproximadamente tomará su aplicación?
¿Qué clase de reactivos tendrá (p. ej., de opción múltiple, de respuesta abierta)?

Origen de las pruebas nuevas

Antes de continuar con el siguiente paso de la elaboración de pruebas, haremos una pausa para considerar esta pregunta: ¿Qué motiva la elaboración de pruebas nuevas? No hay una lista sencilla y definitiva de las motivaciones que están detrás de los proyectos de elaboración de pruebas. Sin embargo, el análisis de las pruebas existentes sugiere *t*res principales motivos de este trabajo.

Primero, muchas de las pruebas que más se usan surgieron en respuesta a alguna *necesidad práctica*. La prueba de inteligencia de Binet, precursora de la Escala de Inteligencia Stanford-Binet, se diseñó para identificar a niños en escuelas parisinas que pudieran necesitar lo que ahora llamamos educación especial. El Stanford-Binet se elaboró para brindar una escala tipo Binet que se pudiera usar con los estadounidenses. La Escala Wechsler Bellevue de Inteligencia, que dio origen a la multitud de escalas Wechsler actuales, se creó para ofrecer una prueba de inteligencia más adecuada para adultos que el Stanford-Binet. Ésos son sólo unos pocos ejemplos de que muchas pruebas surgieron como respuesta a una necesidad muy práctica.

Algunas pruebas se diseñaron a partir de una *base teórica*; por ejemplo el Test de Apercepción Temática (TAT) pretendía ofrecer una medida de la teoría de la personalidad de Murray. La prueba *Primary Mental Abilities* [Capacidades Mentales Primarias] de Thurstone, el prototipo de muchas pruebas multifactoriales de inteligencia posteriores, se diseñó para apoyar la teoría del propio Thurstone sobre las inteligencias múltiples.

Por último, una gran parte del trabajo de elaboración de pruebas se dedica a revisar o adaptar las pruebas existentes; por ejemplo, cada una de las principales baterías de aprovechamiento tiene una nueva edición cada 5 o 10 años. Pruebas como el SAT o ACT son objeto de revisiones más o menos continuas. Las nuevas ediciones de pruebas como las escalas Wechsler y las pruebas de personalidad más populares aparecen de manera regular. Otro tipo de revisión busca modificar una prueba para emplearla con poblaciones especiales.

Preparación de reactivos

La preparación de reactivos incluye su redacción y revisión. No se debe proceder con la redacción de reactivos hasta que el propósito de la prueba esté bien definido y las consideraciones preliminares del diseño se hayan explorado de manera minuciosa. Suponiendo que estos dos primeros pasos se han realizado satisfactoriamente, la preparación de reactivos puede comenzar. Puede ser útil empezar esta sección preguntando: ¿qué es con exactitud un reactivo? Un reactivo consta de cuatro partes (véase figura 6-2). Primero, hay un estímulo al que el examinado responde. Segundo, hay un *formato de respuesta* o método. Tercero, hay *condiciones* que regulan el modo en que se emite la respuesta al estímulo. Cuarto, hay procedimientos para calificar la respuesta, a veces llamados rúbricas de calificación. Describamos brevemente cada uno de estos componentes.

Figura 6-2. Anatomía del reactivo de una prueba.

El estímulo, a menudo llamado **tronco del reactivo**, puede ser una pregunta como las que aparecen en el cuadro 6-2. La primera es de una prueba de inteligencia; la segunda, de una prueba de aprovechamiento; la tercera, de una encuesta de actitudes; y la cuarta, de un inventario de personalidad. El estímulo también puede ser una imagen acompañada de una pregunta oral; por ejemplo, en el Rorschach se presenta una imagen junto con una pregunta acerca de qué ve el examinado. El estímulo también puede ser un aparato como un dinamómetro de mano, pero el "reactivo" no está completo sin una instrucción como "Tómalo con tu mano derecha y apriétalo tan fuerte como puedas".

El formato de respuesta incluye factores como si el reactivo es de opción múltiple o de respuesta abierta. Por ejemplo, cualquiera de los estímulos del cuadro 6-2 podría tener un conjunto de opciones o podría requerir una respuesta abierta. Tratamos varios formatos de respuesta con más detalles en la siguiente sección.

Cuadro 6-2. Ejemplos de estímulos que conforman los reactivos de una prueba

¿Qué significa "pródigo"?
Encuentra el valor de x: Si $6x + 10 = 14$, $x =$ _____
¿Te gusta conocer personas nuevas?
Termina esta oración: Hoy me siento especialmente _____

Quizá no tan evidente como los primeros dos componentes del reactivo, el tercero es decisivo para entender su naturaleza. Las *condiciones que regulan la respuesta* incluyen factores como la existencia de un límite de tiempo para responder, la posibilidad de que el aplicador explore respuestas ambiguas y la manera exacta en que se registra la respuesta, por ejemplo, en una hoja de respuestas o en un cuadernillo de la prueba.

Por último, el procedimiento de calificación es parte crucial del reactivo; en el caso de una prueba de capacidad o de aprovechamiento de opción múltiple, cada reactivo podría calificarse como correcto o incorrecto. Otra alternativa es que se pueda conceder crédito parcial por elegir ciertas opciones. En el caso de los reactivos de respuesta abierta en algunas partes de la Escala Wechsler de Inteligencia para Adultos, una respuesta muy buena merece dos puntos, una respuesta aceptable –pero no particularmente buena– se califica con un punto, mientras que una respuesta incorrecta no recibe ningún punto. Los procedimientos para calificar las respuestas en las técnicas proyectivas pueden ser muy elaborados. Así, los procedimientos de calificación deben ser especificados y comprendidos cuando se consideran los reactivos de una prueba.

Tipos de reactivos

Hay una gran variedad de formas que pueden adoptar los reactivos de una prueba; por lo general, se clasifican en términos del formato de respuesta, el segundo componente de un reactivo considerado con anterioridad. En un nivel muy general, los reactivos se pueden clasificar como de respuesta cerrada o de respuesta abierta.[2] Aquí presentamos sólo los ejemplos más comunes de estos formatos con un breve comentario acerca de sus aplicaciones usuales y sus fortalezas y debilidades.

Reactivos de respuesta cerrada

En los **reactivos de respuesta cerrada** se le presenta al examinado por lo menos dos opciones, pero no más de un número razonable, para que elija una respuesta. Estos reactivos también se denominan de respuesta múltiple, de opción múltiple o de opciones forzadas.

Entre las pruebas más utilizadas, el formato de respuesta cerrada es el más popular. La mayoría de las pruebas de capacidad y de aprovechamiento de aplicación grupal tiene el formato de opción múltiple con cuatro o cinco opciones en cada reactivo. Sin duda, todos los lectores están familiarizados con este tipo de reactivos. Un caso especial del reactivo de opción múltiple es el reactivo de verdadero-falso, que en realidad es de opción múltiple pero con sólo dos opciones: verdadero o falso. El cuadro 6-3 ilustra los reactivos de opción múltiple y de verdadero-falso en pruebas de aprovechamiento.

Cuadro 6-3. Ejemplos de reactivos de opción múltiple y de verdadero-falso de una prueba de aprovechamiento

Reactivo de opción múltiple					
¿Cuál de éstos es un método para determinar la confiabilidad de una prueba?					
		A. Test-retest	B. Estanina	C. Validez	D. Criterio
Reactivo verdadero-falso					
V	F	La estanina es un método para determinar la confiabilidad de una prueba.			

Los formatos de respuesta cerrada son más usuales en el campo de la evaluación de capacidades y aprovechamiento. Sin embargo, también son los que más se usan en las pruebas de personalidad, intereses y actitudes. El cuadro 6-4 muestra los reactivos de opción múltiple y de verdadero-falso en inventarios de intereses y de personalidad.

Cuadro 6-4. Ejemplo de reactivos de opción múltiple y de verdadero-falso de pruebas de intereses y de personalidad

Reactivos de opción múltiple

En cada reactivo, marca si la actividad te gusta (G), te disgusta (D) o no estás seguro (?).				
	G	?	D	
Trabajar con números	O	O	O	
Resolver problemas de resta	O	O	O	

Reactivos de verdadero-falso

En cada reactivo, marca si es verdadero (o casi siempre verdadero) o falso (o casi siempre falso) para ti.

V	F	Me siento deprimido la mayor parte del tiempo.
V	F	Me ha ido muy bien últimamente.

Un caso especial de un formato de respuesta cerrada usado en muchas medidas de actitud es el **formato Likert**.[3] El cuadro 6-5 ilustra este formato. Estos reactivos emplean la escala de cinco puntos, que va desde Completamente de acuerdo hasta Completamente en desacuerdo. Una prueba puede usar una escala de tres, nueve o cualquier número finito de puntos. En una variante, las respuestas se pueden marcar a lo largo de un continuo entre dos polos; luego, las marcas se convierten en una forma numérica. Este procedimiento, a veces llamado **escala de valoración gráfica** (Guilford, 1954) o escala análoga visual (Barker, Pistrang, & Elliott, 1994), se muestra en la figura 6-3. El examinado puede marcar en cualquier punto de la línea; después, las respuestas se convierten en una forma numérica (1-10 en este ejemplo) utilizando la escala que se muestra. Una aplicación interesante de este formato de respuesta es el **diferencial semántico**, en el que un objeto (p. ej., idea, persona u organización) se valora con una serie de escalas en cuyos extremos se encuentran adjetivos opuestos como "duro-suave", "hostil-amigable", "frío-caluroso" y "competente-incompetente".[4] La figura 6-4 ilustra este método.

Cuadro 6-5. Ejemplo del formato Likert en reactivos de actitud

CA = Completamente de acuerdo A = De acuerdo ? = Indeciso D = En desacuerdo CD = Completamente en desacuerdo

	CA	A	?	D	CD
Me encanta el álgebra.	O	O	O	O	O
La raíz cuadrada es genial.	O	O	O	O	O
Me muero de ganas por tomar estadística.	O	O	O	O	O
Los problemas de aritmética son divertidos.	O	O	O	O	O
Me gustaría aprender geometría.	O	O	O	O	O

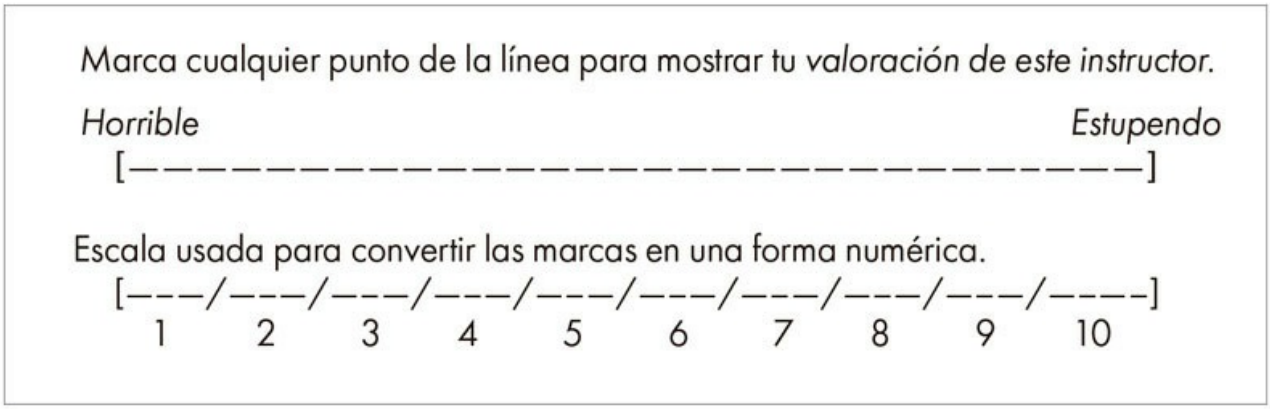


Figura 6-3. Ilustración de una escala de valoración gráfica.

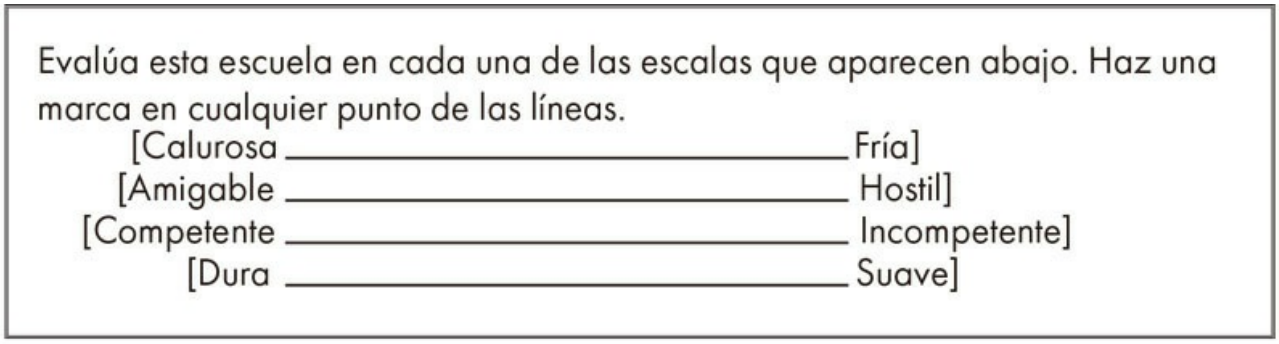


Figura 6-4. Ejemplo del método de diferencial semántico.

Calificación de los reactivos de respuesta cerrada

En el campo de las pruebas de capacidad y de aprovechamiento, la mayoría de los reactivos de respuesta cerrada se califica simplemente como correcto o incorrecto: se concede un punto por una respuesta correcta y cero puntos por una respuesta incorrecta. Luego, la puntuación de la prueba es el número total de respuestas correctas. Sin embargo, hay otras maneras de calificar estos reactivos. Una variante implica conceder crédito parcial por elegir una opción que no es la mejor posible, pero tampoco es un claro error. Otra variante implica dar un peso extra a los reactivos de especial importancia; por ejemplo, con el fin de calcular la puntuación total de la prueba, las respuestas correctas a ciertos reactivos podrían tener un valor de tres puntos, dos puntos las de otros y un punto las de los reactivos restantes. Una variante más para calificar estos reactivos implica utilizar la corrección para las respuestas adivinadas (véase capítulo 3, 49a»). Se han hecho muchas investigaciones en las que se compara la simple calificación de 0 o 1 con estos métodos más complicados. La pregunta es si los sistemas de calificación más complicados proporcionan puntuaciones más confiables o más válidas que la calificación más simple de 0 y 1. La respuesta consta de dos partes. Primero, los sistemas más

complicados suelen proporcionar puntuaciones ligeramente mejores (más confiables o más válidas). Segundo, los autores difieren; algunos dicen que con estas mejoras marginales no vale la pena molestarse con sistemas más complicados (véase, p. ej., Nunnally & Bernstein, 1994). Otros dicen que, con ayuda de la calificación moderna por computadora, esas llamadas complicaciones son triviales, así que cualquier aumento en la confiabilidad o validez es valioso. En Millman y Greene (1993) se puede encontrar una descripción de algunos de estos procedimientos para dar distintos valores a los reactivos o las opciones.

La calificación de reactivos de respuesta cerrada de las pruebas de personalidad, intereses y actitudes se hace de diversas maneras. En algunas aplicaciones, cada reactivo se califica con 1 o 0 de manera análoga al procedimiento de correcta-incorrecta de las pruebas de capacidad. Sin embargo, en este caso, la calificación 1 o 0 no implica que las respuestas sean correctas o incorrectas, sino que una respuesta tiene cierta dirección; por ejemplo, más ansioso, más deprimido, mejor adaptado o más interesado en alguna actividad.

Con frecuencia, se asignan números variables a las diferentes respuestas en reactivos de pruebas de personalidad, intereses y actitudes. Por ejemplo, en la escala Me gusta-?-Me disgusta, podríamos asignar las calificaciones de 3, 2 y 1 o de +1, 0 y −1 a las distintas respuestas, mientras que en la escala de cinco puntos que va de Totalmente de acuerdo a Totalmente en desacuerdo, podríamos asignar 5, 4, 3, 2 y 1 punto o +2, +1, 0, −1 y −2 puntos a las distintas respuestas. Podemos notar que sería posible asignar una puntuación de 1 a las respuestas de Totalmente de acuerdo o De acuerdo y de 0 a todas las demás. El método para calificar estos reactivos se determina en la fase de elaboración de la prueba que corresponde a las consideraciones preliminares del diseño.

Reactivos de respuesta abierta

Los **reactivos de respuesta abierta** presentan un estímulo, pero no limitan al examinado a elegir de un conjunto predeterminado de respuestas. El examinado debe crear o construir una respuesta; *de respuesta libre* es otra denominación común de este formato. Aunque la respuesta del examinado es "libre" en el sentido de que no hay alternativas preexistentes, sí hay condiciones que regulan la respuesta. Por ejemplo, las respuestas se deben dar de manera oral dentro de cierto periodo o tienen que escribirse en forma de ensayo. Por lo común, las pruebas de inteligencia de aplicación individual recurren a un formato de respuesta abierta; por ejemplo, se puede preguntar a un examinado "¿Qué significa pródigo?" o "Si un lápiz cuesta $15 y Jim compra 5 lápices, ¿cuánto pagó?" En cada ejemplo, el examinado construye una respuesta "partiendo de cero". La respuesta puede darse de manera oral o escrita. Una versión muy sencilla de este tipo de reactivos es el formato de *llenar el espacio en blanco*; aquí se omite la palabra o frase clave de una oración. El examinado debe agregar la palabra o palabras faltantes; por ejemplo: Las estaninas son un tipo de puntuación _____. Podemos notar que, en preguntas como ésta, sería fácil usar exactamente el mismo reactivo en un formato de respuesta cerrada.

Uno de los ejemplos más conocidos del formato de respuesta abierta es la **prueba de ensayo**. El reactivo presenta una situación o tema y el examinado escribe una respuesta que puede ir desde unas pocas oraciones hasta varias páginas de texto. La prueba de ensayo podría considerarse como un ejemplo de la categoría más general de **evaluación del desempeño**, en la cual el estímulo se supone que es una situación realista como un problema científico, una tarea en la biblioteca o una producción artística. La respuesta implica resolver el problema, realizar la tarea o producir una obra de arte. En la actualidad las evaluaciones del desempeño atraen la atención en el área educativa como una alternativa a las medidas de opción múltiple de aprovechamiento. Una aplicación popular de la evaluación del desempeño es el uso de portafolios, que, como lo sugiere su nombre, es en esencia una colección de la obra de una persona. Puede crearse un portafolio para los trabajos escritos, los proyectos del laboratorio de ciencias o los análisis estadísticos terminados. Esta colección podría llevarse algunos meses o años. Como otras evaluaciones del desempeño, los contenidos del portafolio se convierten en una herramienta de evaluación cuando su calidad se juzga en alguna dimensión pertinente. En McMillan (2013) pueden encontrarse ejemplos de varias evaluaciones del desempeño.

El formato de respuesta abierta se usa mucho en la evaluación de la personalidad. El cuadro 6-6 muestra dos ejemplos relativamente sencillos de dichos reactivos. Desde luego, los ejemplos clásicos de las medidas de respuesta abierta de personalidad son las técnicas proyectivas como el Rorschach y el TAT. En estas pruebas, cada reactivo presenta un estímulo ambiguo y el examinado tiene la amplia libertad de crear una respuesta. Consideraremos con mayor detalle estas técnicas en el capítulo 14.

Cuadro 6-6. Ejemplos de reactivos sencillos de respuesta abierta que miden personalidad

Asociación de palabras
Diré una palabra y usted dirá la primera palabra que le venga a la mente.
Caliente...
Escuela...
Verano...
Madre...
Frases incompletas
Termine cada oración.
Mi juego favorito es...
Las familias son...
El problema más grande es...

Algunas pruebas conductuales también pueden clasificarse como de reactivos de respuesta abierta; por ejemplo, la técnica del grupo sin líder y la técnica de la canasta son reactivos de respuesta abierta. En el capítulo 13 se puede encontrar una descripción más detallada de estas técnicas.

Calificación de los reactivos de respuesta abierta

Calificar los reactivos de respuesta abierta presenta desafíos especiales justo porque las respuestas pueden ser muy diversas. En casi todos los casos, la calificación requiere del juicio.

Hay dos factores clave para desarrollar puntuaciones útiles de los reactivos de respuesta abierta. El primero es asegurar la confiabilidad interjueces y el segundo es conceptualizar un esquema de calificación.

Ya que calificar reactivos de respuesta abierta requiere, por lo general, de un criterio, el grado de acuerdo entre los jueces (quienes emplean su criterio) es decisivo. Si hay un inadecuado acuerdo entre ellos, las puntuaciones que resulten de la prueba no tendrán sentido. Desde luego, la confiabilidad interjueces no garantiza la validez de las puntuaciones, ni otros tipos de confiabilidad lo hacen, como la de test-retest. Sin embargo, sin esta confiabilidad, todo lo demás está perdido. El punto aquí es que la confiabilidad interjueces es una preocupación especial para los reactivos de respuesta abierta.

Conceptualizar un esquema para calificar los reactivos es un reto mayor, pues los tipos de esquemas varían ampliamente. No parece factible proporcionar una lista exhaustiva, así que mejor daremos ejemplos de algunos de los métodos que se han desarrollado; primero consideraremos algunos ejemplos del campo de las pruebas de aprovechamiento y luego dirigiremos nuestra atención a las de personalidad.

Resumen de puntos clave 6-2

Algunos métodos para calificar ensayos y otras producciones

- Holístico
- Analítico
- Sistema de puntos
- Calificación automatizada

Varios métodos diferentes han surgido para calificar los ensayos. Una distinción común es entre la calificación **holística** y la **analítica.** En la **calificación holística**, el lector –es decir, la persona que califica el ensayo– hace un juicio sencillo, general, holístico acerca de la calidad del ensayo. La puntuación asignada al ensayo refleja su juicio general; la escala puede tener cualquier número de puntos, por ejemplo, 1-4, 1-10 o 1-100. La lectura se suele llevar a cabo de manera rápida, sin hacer correcciones ni notas en el papel. ¿Qué calidad se juzga que tiene el ensayo? Eso depende de la aplicación específica.

El ensayo puede valorarse en términos de la calidad de la expresión escrita, en el caso de una prueba de composición en inglés, o del conocimiento de un tema, en el caso de una prueba de historia. La característica clave de la calificación holística es que sólo hay

una puntuación global basada en la calidad total del ensayo.

En la **calificación analítica**, el mismo ensayo (u otra producción) se valora en diversas dimensiones. Requiere una especificación previa de las dimensiones importantes de la calidad del ensayo. El mismo juez puede realizar las valoraciones separadas, o diferentes jueces, uno para cada dimensión, pueden hacerlo. En el ensayo de composición en inglés puede valorarse de manera independiente a) la corrección gramatical, b) la organización y c) el uso del vocabulario. El ensayo de historia podría valorarse de manera independiente de acuerdo con el uso de los hechos históricos, la identificación de temas principales y la calidad de la escritura. Desde la perspectiva de la medición, la calificación analítica supone obviamente que hay cierta independencia significativa entre los rasgos especificados en el esquema de calificación analítica. A menudo, tal independencia parece estar ausente, como lo indican las correlaciones extremadamente altas entre las escalas de varios esquemas analíticos. Debe establecerse cierto grado de independencia entre las escalas antes de adoptar el sistema de calificación analítica.

Un tercer método para calificar un ensayo es el **sistema de puntos**. Aquí, ciertos puntos tienen que incluirse en una respuesta "perfecta". El juez sólo determina la presencia o ausencia de cada punto. El ejemplo más sencillo de este sistema es una prueba de memoria pura, por ejemplo "Di los diez mandamientos". Se otorga un punto por cada mandamiento; desde luego, incluso en este sistema se requiere el criterio del juez, excepto en los casos más triviales. ¿"Irás a la iglesia los domingos" se toma como respuesta correcta del tercer mandamiento? ¿Los mandamientos deben decirse en el orden tradicional?

Los diferentes métodos para calificar ensayos también pueden aplicarse a una gran variedad de *valoraciones de producciones*; de hecho, los ensayos son sólo un tipo de producción. Estos métodos pueden aplicarse a la evaluación del desempeño de producciones artísticas, proyectos científicos, habilidad para hablar en público y muchas otras. Al calificar un portafolio, se debe tomar una decisión no sólo acerca del método de calificación, sino también de la característica del propio portafolio. Se pueden calificar todas las entradas del portafolio, sólo las mejores o la cantidad de progreso que se muestra de las primeras a las últimas entradas.

Hemos hecho hincapié varias veces en que la calificación de reactivos de respuesta abierta requiere del juicio. Esto se lleva tiempo, es costoso y puede estar lleno de problemas relacionados con la confiabilidad de los jueces. Los investigadores ahora estudian la aplicación de sistemas expertos de cómputo –llamada **calificación automatizada**– para calificar los reactivos de respuesta abierta. No se debe confundir la calificación automatizada con la simple calificación mecánica de respuestas a reactivos de opción múltiple donde se llena un espacio en blanco. La calificación automatizada, como el término ha evolucionado en la literatura de las investigaciones, implica el desarrollo de programas sofisticados de cómputo que simulan el proceso de aplicar el criterio humano a los reactivos de respuesta libre. Por ejemplo, un proyecto aplicó sistemas de calificación automatizada a la evaluación del desempeño de las habilidades de un médico para manejar a los pacientes (Clauser, Swanson, & Clyman, 1999). En otro proyecto se

aplicó un sistema de calificación automatizada en la evaluación de las respuestas de arquitectos a problemas de respuesta abierta relacionados con la arquitectura (Williamson, Bejar, & Hone, 1999). Uno de los primeros trabajos en esta línea fue el de las puntuaciones generadas por computadora de Ellis Page para calificar la calidad con que estaba escrito un ensayo. Wresch (1993) ofrece un repaso histórico de estos esfuerzos; en Page y Petersen (1995) se encuentra un repaso semi-popular del trabajo actual de Page con el Project Essay Grade (PEG [Proyecto Valoración de Ensayos]). La cuestión clave de estos proyectos es si el sistema automatizado se aproxima al juicio humano experto. El *Graduate Management Admissions Test* [Prueba de admisión para graduados en gestión de empresas] ahora usa una computadora para generar una de las puntuaciones para la parte del ensayo de la prueba y el *Graduate Record Exam* [Examen para graduados] anunció su intención de empezar a hacerlo. Es probable que en el futuro veamos una rápida expansión del uso de varios sistemas de calificación automatizada para los reactivos de respuesta abierta. En Dikli (2006), Drasgow, Luecht y Bennett (2004), Shermis y Burstein (2003) y Shermis y Daniels (2003) se pueden encontrar resúmenes de la aplicación de la calificación automatizada, especialmente de ensayos.

En el campo de las pruebas de personalidad, las técnicas proyectivas ofrecen un ejemplo clásico de los reactivos de respuesta abierta. Consideraremos las técnicas proyectivas de una manera más sistemática en el capítulo 14. Aquí sólo ilustramos algunas de estas técnicas en cuanto a su calificación.

Los métodos comunes para calificar el Rorschach se apoyan en la especificación de categorías y en el conteo del número de respuestas que caen en ellas. El cuadro 6-7 muestra una categoría que se usa bastante: localización, es decir, la localización de una tarjeta usada como punto de referencia para la respuesta. El juez clasifica las respuestas de cada tarjeta de acuerdo con estas categorías. Aquí, el esquema conceptual consiste en a) una lista de categorías y b) la noción de contar los enunciados.

Cuadro 6-7. Categorías de muestra para calificar las respuestas de "localización" a una lámina del Rorschach

Determina la localización en una tarjeta usada como punto de referencia para la respuesta:	
Completa	La mancha de tinta completa usada para formular la respuesta
Detalle común	Una parte bien definida que se observa comúnmente
Detalle inusual	Se usa una parte inusual
Espacio	La respuesta está definida por un espacio en blanco

El *Rotter Incomplete Sentences Blank* (RISB [Frases incompletas de Rotter]; Rotter, Lah, & Rafferty, 1992) ofrece otro ejemplo de un esquema conceptual para calificar una prueba proyectiva. El Rotter consiste en 40 oraciones incompletas similares a las que aparecen en el cuadro 6-6. Cada respuesta se califica en una escala de seis puntos de acuerdo con el grado de inadaptación manifestada. Las valoraciones de los 40 reactivos se suman para obtener una puntuación total de adaptación. El manual del RISB contiene instrucciones específicas en relación con los indicadores de adaptación/inadaptación. Así,

el esquema conceptual es mirar las respuestas en términos de los indicadores de adaptación, valorarlas en una escala numérica simple y luego sumar estas valoraciones para obtener una puntuación total. En el capítulo 14 se habla más sobre el Rorschach y el RISB.

¡Inténtalo!

El RISB se califica de acuerdo con el grado de inadaptación que indican las respuestas. ¿Puedes pensar en otro esquema conceptual que pueda usarse para calificar las respuestas?

El manual de la prueba tiene un papel esencial para asegurar que los reactivos de respuesta abierta produzcan puntuaciones significativas; para ello, debe especificar qué tipo de capacitación se necesita para calificar los reactivos. El manual también debe explicar los fundamentos para calificarlos y poner ejemplos. Las instrucciones para calificar reactivos de respuesta abierta, por lo general con ejemplos de respuestas de distintos niveles, a menudo se denominan *rúbricas* de calificación. El manual también debe informar los resultados de estudios de confiabilidad de interjueces.

Ventajas y desventajas de los reactivos de respuesta cerrada frente a los de respuesta abierta

En la literatura psicométrica y en los medios públicos se defienden con vehemencia las ventajas relativas de los reactivos de respuesta abierta y los de respuesta cerrada. Tratemos de resumir los puntos principales de estas discusiones.

Los reactivos de respuesta cerrada tienen tres ventajas principales. La primera es la confiabilidad de su calificación; ya que no se requiere emplear el criterio, o se necesita poco de él, una fuente importante de varianza no confiable se elimina. La confiabilidad interjueces es, en esencia, perfecta para los reactivos de respuesta cerrada; en cambio, para los reactivos de respuesta abierta es un problema considerable. Esta preocupación por la confiabilidad fue el estímulo para desarrollar las primeras versiones de opción múltiple de las pruebas de aprovechamiento a principios del siglo XX. Las pruebas de opción múltiple no se desarrollaron para acomodarse a la calificación por medio de máquinas, como a menudo se piensa. De hecho, no había máquinas para calificarlas en ese tiempo. El reactivo de opción múltiple se convirtió en el formato preferido porque proporcionaba puntuaciones más confiables que los reactivos de respuesta abierta. En Ebel (1979) se puede encontrar un repaso histórico de los factores que llevaron al desarrollo de las primeras medidas de aprovechamiento de opción múltiple.

La segunda ventaja importante de los reactivos de respuesta cerrada es la eficiencia temporal; en cierta cantidad de tiempo, un examinado puede, por lo general, terminar más reactivos de respuesta cerrada que de respuesta abierta. Por ejemplo, en 20 minutos un examinado podría terminar con facilidad 30 reactivos de vocabulario de opción

múltiple, mientras que en el mismo tiempo, el examinado podría terminar sólo 10 reactivos de vocabulario en el formato de respuesta abierta. En el caso de una medida de aprovechamiento, en 20 minutos una persona podría terminar un ensayo en vez de responder 30 reactivos de opción múltiple. Ya que, por lo general, la confiabilidad aumenta en función del número de reactivos, esta segunda ventaja, como la primera, se reduce a una cuestión de confiabilidad. Y estas ventajas también se relacionan con la validez debido a la relación entre confiabilidad y validez.

La tercera ventaja de los reactivos de respuesta cerrada es la eficiencia en la calificación. Un empleado de oficina o un escáner electrónico pueden calificar estos reactivos con mucha rapidez. Esta ventaja fue el principal estímulo para desarrollar las primeras pruebas de capacidad y personalidad de aplicación grupal durante la Primera Guerra Mundial. Podemos notar que esta ventaja es independiente en términos lógicos de la confiabilidad interjueces.

A menudo, se citan tres ventajas principales de los reactivos de respuesta abierta. Primero, permiten hacer con mayor facilidad una observación de la conducta y los procesos al responder la prueba. De alguna manera, esta ventaja se relaciona más con el modo de aplicación (individual, en vez de grupal) que con el formato de la respuesta. Sin embargo, el formato de respuesta abierta facilita la observación de la motivación del examinado, ansiedad o formas de acercarse a los problemas, de modo que los reactivos de respuesta cerrada no lo harían aunque se aplicaran de manera individual.

La segunda ventaja del formato de respuesta abierta, sobre todo en el campo de la medición de la personalidad, es que permite explorar áreas inusuales que podrían nunca salir a la luz por medio del formato de respuesta cerrada. Desde luego, la pertinencia de este argumento depende de la exhaustividad de la prueba de respuestas cerradas; si en realidad es exhaustiva, debería, por definición, sacar a la luz toda la información importante. La pregunta es si esas medidas son, en efecto, exhaustivas en su medición de la personalidad.

En el campo de las pruebas de aprovechamiento, algunos autores creen que el tipo de reactivos influye en el desarrollo de los hábitos de estudio de los alumnos. De manera más específica, hay una sensación de que usar reactivos de opción múltiple promueve el aprendizaje de memoria y un enfoque atomista de las materias y temas de estudio, mientras que los reactivos de respuesta abierta promueven un método de estudio más holístico y significativo.

En Hogan (2013), Traub, (1993) y Rodriguez (2002, 2003) se encuentran datos de las investigaciones realizadas sobre los reactivos de respuesta cerrada en comparación con los de respuesta abierta, en especial de pruebas de capacidad y aprovechamiento. Los desarrollos actuales en los sistemas de calificación automatizada pueden influir de manera significativa en la evaluación futura de las ventajas y desventajas relativas de los reactivos de respuesta cerrada frente a los de respuesta abierta.

Sugerencias para escribir reactivos de respuesta cerrada

Hay numerosas listas de sugerencias para redactar reactivos de respuesta cerrada, sobre todo de opción múltiple para las pruebas de aprovechamiento. Haladyna y Downing (1989a, 1989b) prepararon una taxonomía de estas así llamadas reglas para redactar reactivos, obtenida de una investigación en 46 libros de texto y fuentes similares. Después, Haladyna (1994, 1999, 2004) dedicó un libro entero a la elaboración de estas reglas y a la investigación relacionada con su validez. La edición actual de su libro es, sin duda, la mejor fuente de consejos para redactar reactivos de respuesta cerrada y para determinar si las sugerencias hacen alguna diferencia en la calidad de los reactivos. La lista actual contiene 31 directrices (36 si cuentas las subpartes de una entrada). Cualquiera que necesite ayuda con la redacción de reactivos de respuesta cerrada, sobre todo si son para pruebas de capacidad o aprovechamiento, deben consultar esta fuente. Entre los ejemplos de las directrices se encuentran cuestiones como evitar el uso de "todas las anteriores" como opción y mantener todas las opciones más o menos de la misma extensión. Una curiosa ramificación de este trabajo, después elaborada por Rodriguez (2005), es la recomendación de que tres opciones es el número óptimo, a pesar de que casi todas las pruebas importantes usan cuatro o cinco opciones. Haladyna (2004, p. 98) señaló que "los escritores de reactivos deben aplicar estas directrices de manera sensata, pero no rígida, ya que la validez de algunas de ellas aún está en entredicho". De hecho, algunos autores concluyen sus listas de reglas con ésta: Hacer caso omiso de cualquiera de estas reglas cuando parezca haber una buena razón para hacerlo.

Nosotros somos reacios a inventar una lista más de reglas para escribir reactivos; sin embargo, nos aventuraremos a opinar que casi todas las reglas existentes se reducen a estas tres: *tener un buen contenido*, *no regalar la respuesta correcta* y *hacerlo de manera sencilla y clara*. Además, las dos primeras están limitadas a las pruebas de capacidad y aprovechamiento, mientras que sólo la última es para las medidas de personalidad, intereses y actitudes.

Sugerencias para redactar reactivos de respuesta abierta

Como indicamos antes, los libros de texto y los artículos están llenos de sugerencias para redactar reactivos de respuesta cerrada, pero son más limitadas en el caso de los reactivos de respuesta abierta. Quizá la naturaleza tan abierta de estos reactivos hace más difícil formular consejos específicos. Es interesante que el primer consejo que dan muchos creadores experimentados de pruebas es tratar de evitar este tipo de reactivos y, en su lugar, usar los de respuesta cerrada, sobre todo por el tema de la confiabilidad interjueces que ya discutimos.

Hogan y Murphy (2007) resumieron las sugerencias que encontraron en 25 fuentes para preparar y calificar reactivos de respuesta abierta. El lector interesado puede consultar esa fuente para conocer la gran cantidad de sugerencias de esos autores. Al igual que con los reactivos de respuesta cerrada, en el caso de estos reactivos los consejos tienden a orientarse hacia las pruebas de capacidad y aprovechamiento.

Mencionamos aquí sólo algunos de los puntos en que hay mayor consenso para dar sabor a las recomendaciones.

1. Asegurarse de que la tarea es clara. Con los reactivos de respuesta cerrada, la tarea del examinado queda clara al mirar las posibles respuestas, mientras que en los reactivos de respuesta abierta esa ayuda no está presente. De ahí que se necesita tener mayor cuidado al formular y clarificar las instrucciones de estos reactivos.

2. Evitar el uso de reactivos opcionales. Los psicómetras fruncen el entrecejo frente a la muy frecuente práctica de proporcionar reactivos opcionales (p. ej., responder 3 de 5 preguntas), pues interfiere en la comparabilidad de la tarea.

3. Ser específico respecto al sistema de calificación mientras se prepara el reactivo. Y, desde luego, usar ese sistema cuando realmente se esté calificando. Una práctica común en el caso de estos reactivos es preparar el reactivo, aplicarlo y suponer que el método de calificación será claro más tarde. La manera en que el reactivo se calificará, de preferencia con respuestas muestra, debe ser clara antes de que se aplique el reactivo. Esta sugerencia se aplica sin importar la generalidad de la respuesta. Es igual de importante al calificar los reactivos en que se llena el espacio en blanco que al calificar ensayos extensos, evaluaciones de desempeño o técnicas proyectivas.

4. Calificar de manera anónima. Es decir, la persona que califica no debe conocer la identidad del examinado. Esto ayuda a centrarse en la respuesta real y evitar el efecto "halo".

5. Cuando hay reactivos múltiples, calificarlos uno *a la vez*. Como en la sugerencia previa, esto ayuda a evitar el efecto "halo", en el que la respuesta a un reactivo influye en la calificación de la respuesta a otro reactivo.

Algunas consideraciones prácticas para redactar reactivos

Aquí consideramos algunas cuestiones prácticas relacionadas con la redacción de reactivos. Primero, si preparamos un conjunto de reactivos de prueba, ¿cuántos de ellos debemos escribir? Esta pregunta no tiene una respuesta definitiva, ya que ésta depende en parte de tomar buenas decisiones en la etapa preliminar del diseño, por ejemplo, acerca del tipo apropiado de reactivos y de la investigación minuciosa del área a la que corresponde la prueba. La respuesta también depende de hacer un trabajo razonable de pruebas informales para asegurarse de que los prototipos de los reactivos que se tienen pensados funcionarán. Con estos requisitos en mente, una regla general común es preparar dos o tres veces tantos reactivos como sean necesarios para la prueba final. Así, si la prueba final tendrá 50 reactivos, se deben preparar 100 o 150 para ponerlos a prueba. Consideremos algunas desviaciones extremas de esta regla general. Si la prueba final tendrá 50 reactivos y sólo se prueban 55, casi sin duda el análisis de reactivos revelará más de 5 reactivos con características indeseables. Con un margen de sólo 5 reactivos, será necesario incluir algunos en la prueba final aunque no sean muy buenos. En el otro extremo, supongamos que se preparan 500 reactivos de prueba. Debe señalarse que preparar 500 reactivos de cualquier tema es una tarea difícil; más

importante, si se seleccionan 50 reactivos de un fondo de 500, es muy probable que se saque provecho de algunos factores fortuitos que no sobrevivirían a un proceso de validación cruzada. (Véase la discusión sobre validación cruzada del capítulo 5). Si es necesario probar 500 reactivos para obtener 50 servibles, es probable que se necesite replantear el enfoque de la prueba.

En proyectos importantes para elaborar pruebas, una vez que se redactan los reactivos, se someten a una revisión desde varias perspectivas. Primero, se revisa su claridad, adecuación gramatical y conformidad con las reglas que mencionamos antes. Segundo, en el caso de las pruebas de aprovechamiento, expertos en el campo del contenido pertinente revisan que éste sea el correcto. En el caso de las pruebas de personalidad, a menudo se pide a los clínicos que revisen la pertinencia de los reactivos.

Tercero, en años recientes se ha vuelto una costumbre tener un panel que revise los reactivos en cuanto a posibles sesgos de género, raciales o étnicos. Discutiremos estas revisiones con mayor detalle cuando tratemos la neutralidad de las pruebas más adelante en este capítulo.

Análisis de reactivos

Uno de los pasos decisivos en la elaboración de pruebas es el **análisis de reactivos**, el cual implica un análisis estadístico de los datos obtenidos en la prueba de los reactivos. Con base en este análisis se eligen los reactivos que formarán parte de la prueba final. Así, lo que aquí hemos llamado análisis de reactivos consiste en realidad en tres procesos estrechamente relacionados: prueba de los reactivos, análisis estadístico y selección de reactivos. Discutiremos cada proceso en esta sección.

¿Por qué es importante el análisis de reactivos? Como señalamos antes, la gran mayoría de pruebas educativas y psicológicas consiste en un conjunto de reactivos individuales, que son como los bloques de una construcción. Controlamos las características de una prueba controlando los reactivos que la conforman. Si queremos una prueba fácil, usamos reactivos fáciles; si queremos una prueba con una confiabilidad de consistencia interna alta, usamos reactivos que tengan correlaciones altas entre sí. El análisis de reactivos es el conjunto de procedimientos que nos permite ejercer este control. Además, ya que las características del reactivo determinan las características importantes de una prueba, los manuales se refieren con frecuencia a los resultados del análisis de reactivos. De ahí que, para ser un lector informado de estos manuales, debamos estar familiarizados con los conceptos y técnicas del análisis de reactivos.

¡Inténtalo!

Accede a la reseña de una prueba en cualquiera de las ediciones del Mental Measurements Yearbook de Buros, sea en formato electrónico o impreso. Revisa la reseña y ve qué se dice acerca del proceso de construcción de la prueba. Pon especial atención a las referencias relacionadas con los estadísticos de los reactivos. ¿Cuáles se mencionan?

Resumen de puntos clave 6-3

Las tres fases del análisis de reactivos

1. Prueba de los reactivos
2. Análisis estadístico
3. Selección de reactivos

Prueba de reactivos

Hay dos etapas de la prueba de reactivos: formal e informal. Los datos del análisis de reactivos se basan en la prueba formal; sin embargo, antes de llevar a cabo la prueba formal, lo habitual y prudente es realizar una prueba informal. Esto se suele hacer con sólo algunos casos, digamos entre cinco y diez parecidos a aquellos a quienes está dirigida la prueba. A menudo, los reactivos ni siquiera se califican de una manera formal, pues se pide a los individuos que participan en la prueba informal que hagan comentarios sobre los reactivos y las instrucciones de la prueba. Se puede pedir a los individuos "pensar en voz alta" mientras responden los reactivos, lo cual puede ser de especial ayuda con los formatos o métodos novedosos, pues permite al creador de la prueba identificar una redacción ambigua, interpretaciones inesperadas de un reactivo, confusión acerca de los métodos para responder y otras anomalías como éstas. La prueba informal puede evitar gastar recursos en la fase de prueba formal. No tiene sentido recoger datos de varios cientos de examinados y realizar elaborados análisis estadísticos con reactivos que ni siquiera comprenden los examinados.

La prueba formal de reactivos implica aplicar los reactivos de la nueva prueba a muestras de examinados, las cuales deben ser representativas de la población a la que está dirigida. Por ejemplo, si la prueba está pensada para usarse con niños normales de 3 a 6 años de edad, la muestra de prueba debe ser representativa de este grupo. Si la prueba está pensada para aspirantes a la universidad, la muestra de prueba debe ser representativa de los aspirantes a la universidad.

Las muestras de prueba de reactivos a menudo no son tan grandes como las que se usan para crear las normas de la prueba; sin embargo, es obvio que necesitan ser lo suficientemente grandes para proporcionar datos estables. Muestras de varios cientos de individuos suelen ser adecuadas cuando se usan los procedimientos clásicos de análisis de reactivos, como los definiremos más adelante, pero usar los procedimientos de la teoría de la respuesta al reactivo puede requerir muestras mucho más grandes.

Los creadores de pruebas usan uno de varios procedimientos diferentes para llevar a cabo la prueba de reactivos, la cual puede implicar un estudio independiente en el que el único propósito es poner a prueba los reactivos. O los reactivos de prueba pueden estar incrustados en la aplicación de una prueba existente que proporciona puntuaciones regulares sobre la prueba, pero los reactivos de prueba no contribuyen a la puntuación. Muchas pruebas de aplicación adaptable por computadora usan este procedimiento. Después de determinar los estadísticos del reactivo, como se describe en la siguiente sección, un reactivo puede estar incluido en la versión "en vivo" de la prueba.

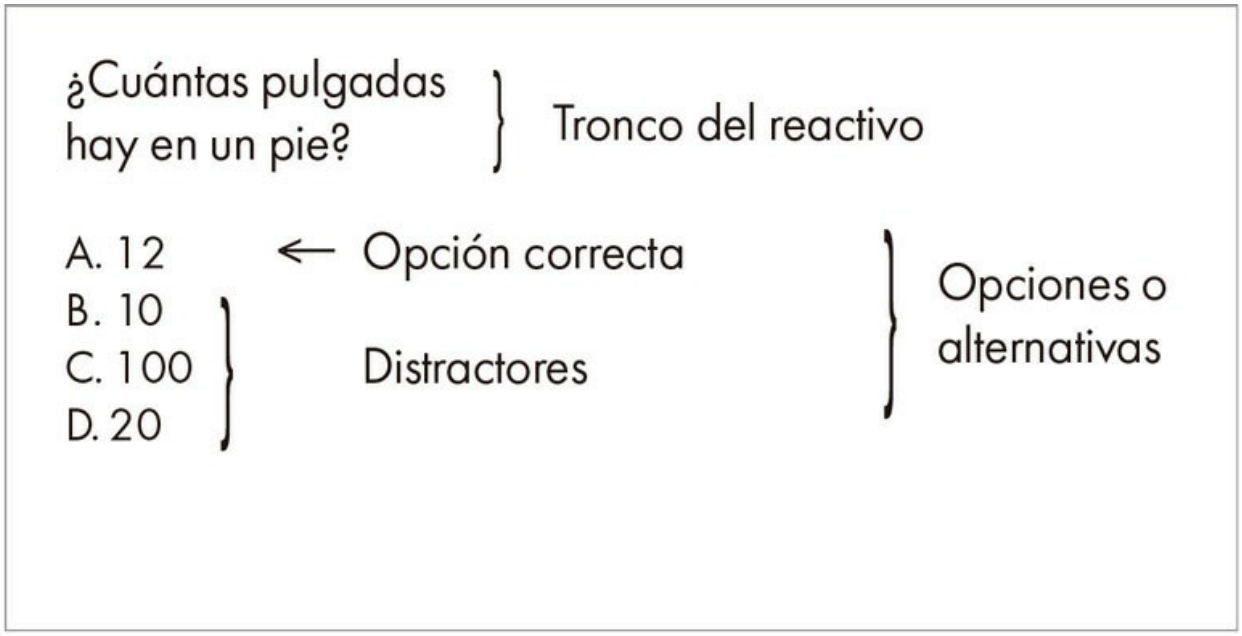


Figura 6-5. Anatomía de los reactivos de opción múltiple.

Estadísticos de los reactivos

Gran parte del vocabulario que se usa en el análisis de reactivos se origina en las aplicaciones de las pruebas de aprovechamiento y de capacidad, en especial las de reactivos de opción múltiple. En estos dominios, hay una opción correcta y varias incorrectas; sin embargo, los procedimientos de análisis de reactivos también trabajan con pruebas de otros dominios, por ejemplo, con pruebas de personalidad o encuestas de actitudes. La terminología desarrollada en el dominio cognitivo se transmite a menudo a otros dominios, aunque su uso a veces es un poco forzado. La figura 6-5 muestra la terminología que se suele usar con dichos reactivos.

Dificultad del reactivo

La prueba formal de reactivos da por resultado un conjunto de estadísticos. Los procedimientos tradicionales de análisis de reactivos, los que se desprenden de la teoría clásica de las pruebas, dependen de dos conceptos: índice de dificultad e índice de discriminación. La **dificultad del reactivo** se refiere al porcentaje de examinados que responden correctamente, si se trata de reactivos que se califican como correcto/incorrecto, o en cierta dirección, si se trata de reactivos para los cuales no hay respuesta correcta, por ejemplo, "de acuerdo" en un reactivo de actitud. En el caso de reactivos calificados con correcto/incorrecto, la "dificultad" es realmente un índice de "facilidad", pues se refiere al porcentaje de respuestas correctas. Sin embargo, el término *dificultad del reactivo* está muy consolidado en la literatura psicométrica.

Los niveles de dificultad suelen denominarse **valores p**, donde p representa el porcentaje o proporción. Así, un reactivo con un valor p de 85 es un reactivo fácil: 85% de los examinados lo respondieron correctamente. Un reactivo con valor p de 25 es muy difícil: sólo 25% de los examinados lo respondieron correctamente.

Discriminación del reactivo «148-150a

La **discriminación del reactivo** se refiere a la capacidad del reactivo para diferenciar en términos estadísticos de la manera deseada entre grupos de examinados. Aquí, el término **discriminación** no hace referencia a una discriminación sociológica o jurídica debido a la raza, género o religión. A causa del potencial malentendido que puede surgir en este punto, haríamos muy bien usando algún otro término para referirnos a esta diferenciación estadística. Sin embargo, como sucede con dificultad del reactivo, discriminación del reactivo es un término muy consolidado en la literatura psicométrica y no es probable que desaparezca.

¿Qué clase de discriminación o diferenciación queremos de un reactivo? Por lo general, queremos que el reactivo pueda diferenciar entre individuos que tienen más del rasgo que

tratamos de medir y los que tienen menos del rasgo. Ya que los reactivos son los bloques que conforman la prueba, los que pueden diferenciar de esta manera harán una buena prueba. Los siguientes parágrafos describen a) cómo definir los grupos con más o menos de un rasgo y b) cómo expresar el grado en que un reactivo puede diferenciarlos.

Para determinar si un reactivo diferencia entre quienes tienen más o menos del rasgo que queremos medir, necesitamos identificar grupos con más o menos de ese rasgo, para lo cual existen dos métodos que se usan con frecuencia. El primero se llama *método externo* y el segundo, *método interno*. El punto de referencia de los términos *externo* e *interno* es la prueba misma; en el método externo, la base para identificar al grupo es externa a la prueba, mientras que en el método interno, la base es interna a la prueba. El método externo depende de tener dos (o más) grupos diferenciados en el rasgo pertinente de acuerdo con algún criterio externo. Consideremos estos dos ejemplos de grupos definidos externamente. Primero, supongamos que estamos elaborando un cuestionario para medir depresión. Tenemos un grupo de 50 individuos diagnosticados con depresión por un equipo de psicólogos y otro grupo de 50 individuos que experimentan reacciones fóbicas moderadas, pero sin otros síntomas clínicos significativos. Queremos reactivos que discriminen o diferencien el grupo de deprimidos del grupo de no deprimidos. Segundo, supongamos que estamos elaborando una prueba de eficiencia en el uso de Microsoft Access, un programa de cómputo para crear bases de datos. Tenemos un grupo de 100 individuos que han terminado una capacitación en Access de tres semanas y otro grupo de individuos que saben usar la computadora en general, pero no han sido capacitados en Access. Esperamos que nuestros reactivos discriminen entre estos dos grupos.

En el método interno para crear grupos con más o menos del rasgo que tratamos de medir, calificamos la prueba entera y, luego, identificamos a los que tuvieron puntuaciones más altas y más bajas. El supuesto es que la prueba entera es una medida razonablemente válida de ese rasgo. Entonces determinamos el grado en que un reactivo individual diferencia entre los que tuvieron puntuaciones altas y bajas. En esencia, determinamos el grado en que el reactivo diferencia entre personas de la misma manera en que lo hace la puntuación total. En el caso de muchos rasgos que deseamos medir, no tenemos un buen indicador externo o es muy difícil conseguirlo; de ahí que el método interno se use con mucha mayor frecuencia que el externo para elaborar una prueba.

En el método interno se emplea una de varias formas para dividir a los individuos en puntuaciones altas y puntuaciones bajas. Empezamos con la distribución de las puntuaciones totales[5] de la prueba. Los **grupos "alto" y "bajo"** pueden identificarse como las mitades, terceras o cuartas partes que se encuentran en el extremo superior e inferior de la distribución, respectivamente. Otra división que se usa con frecuencia es la de 27% superior e inferior.[6] Para tener un análisis completo, cuando se usan 25%, 27% o 33% superior e inferior, es habitual también determinar el desempeño de los grupos intermedios aunque no se tomen en cuenta en el análisis de discriminación. (Para obtener el índice de dificultad, sí se toman en cuenta todos los casos.)

Mientras la dificultad del reactivo casi siempre tiene sólo un indicador universal –el

valor p–, la discriminación del reactivo puede expresarse de varias maneras diferentes. Lo más común es representar el grado de discriminación mediante D (que significa diferencia o discriminación) o *r* (la correlación entre el desempeño en el reactivo y el criterio externo o la puntuación total de la prueba). Por lo común, D se define como la simple diferencia en el porcentaje de respuestas correctas en los grupos "superior" e "inferior". En la práctica, encontramos varios tipos diferentes de coeficientes de correlación (*r*) para expresar las relaciones reactivo-prueba o reactivo-criterio. El tipo de *r* depende de ciertos supuestos que se hacen en casos particulares acerca de la naturaleza de las variables implicadas; los tipos más comunes incluyen las correlaciones *r* biserial (r_{bis}) y *r* biserial puntual (r_{bisp}). En la literatura psicométrica, también encontraremos referencias a la correlación tetracórica (r_{tet}), al coeficiente phi (ϕ) y al término *r corregida*, el cual se usa cuando la correlación entre el reactivo y la prueba se basa en una puntuación total que excluye el reactivo que se está analizando. Todos los métodos proporcionan casi la misma información acerca del poder de discriminación de un reactivo; además de encontrar varias maneras de determinar la discriminación del reactivo, encontramos varias maneras de nombrar el índice. Sin importar el método específico que se use, el resultado puede llamarse índice de discriminación, correlación reactivo-total o índice de validez del reactivo.

El cuadro 6-8 muestra un conjunto de datos empleados en un análisis de reactivos de una prueba muy sencilla de 10 reactivos. Primero, las pruebas se califican y, luego, los casos se ordenan de mayor a menor. El valor p y el índice de discriminación derivan de las respuestas a los reactivos, proceso bastante tedioso para hacerlo a mano, por lo que prácticamente siempre se hace con ayuda de algún programa diseñado específicamente para el análisis de reactivos.

Cuadro 6-8. Ejemplo de datos ordenados para el análisis de reactivos

Reactivos (1 = correcto, 0 = incorrecto)												
Caso	Puntuación	1	2	3	4	5	6	7	8	9	10	
1	10	1	1	1	1	1	1	1	1	1	1	
2	9	1	1	1	1	1	1	1	1	0	1	
3	9	1	1	1	1	1	1	1	0	1	1	Grupo alto
4	8	1	1	1	1	1	1	0	1	0	1	
5	8	1	1	1	1	1	1	1	0	0	1	
6	8	1	1	1	1	1	1	0	1	0	1	
-												
-												
-												
-												
-												
95	3	1	0	1	1	0	0	0	0	0	0	
96	3	1	1	1	0	0	0	0	0	0	0	
97	3	1	0	1	1	0	0	0	0	0	0	
98	2	1	0	1	0	0	0	0	0	0	0	Grupo bajo
99	2	1	1	0	0	0	0	0	0	0	0	
100	2	1	0	1	0	0	0	0	0	0	0	

Ejemplos de los estadísticos del reactivo

El cuadro 6-9 presenta datos de cinco reactivos de una prueba de aprovechamiento. Examinemos estos datos para ilustrar lo que se puede llegar a saber a partir de un análisis de reactivos. La columna de la izquierda, titulada "Reactivo", presenta el número de reactivos. Hay tres entradas en la columna "Estadísticos del reactivo": Prop. correcta (proporción de respuestas correctas a un reactivo), Índice de disc. (índice de discriminación del reactivo), Biser. puntual (coeficiente de correlación biserial puntual entre el desempeño en este reactivo y la puntuación total de la prueba de 27% de los casos superiores e inferiores).

Cuadro 6-9. Datos muestra del análisis de reactivos de una prueba de aprovechamiento[a]

Reactivo	Estadísticos del reactivo			Estadísticos de las alternativas				
					Prop. de aprob.			
	Prop. correcta	Índice de disc.	Biser. puntual	Alt.	Total	Inferior	Superior	Clave
6	.56	.50	.43	1	.56	.36	.87	*
				2	.26	.45	.07	
				3	.10	.09	.07	
				4	.05	.00	.00	

10	.62	.10	.04	1	.05	.00	.00	
				2	.62	.64	.73	*
				3	.00	.00	.00	
				4	.31	.36	.27	
23	.26	.40	.37	1	.03	.09	.00	
				2	.08	.18	.00	
				3	.26	.00	.40	*
				4	.56	.55	.60	
28	.97	.09	.24	1	.00	.00	.00	
				2	.03	.09	.00	
				3	.00	.00	.00	
				4	.97	.91	1.00	*
29	.69	.05	.03	1	.69	.55	.60	*
				2	.08	.09	.13	
				3	.15	.27	.20	
				4	.08	.09	.07	

[a] Formato adaptado de ITEMAN ™, un componente del paquete de análisis de pruebas y reactivos diseñado por Assessment Systems Corporation, reproducido con permiso.

En "Estadísticos de las alternativas", encontramos las siguientes entradas: Alt. (alternativa u opción; en esta prueba, cada reactivo tiene cuatro opciones); Prop. de aprob. (proporción de aprobación o elección de cada opción) en cada uno de los siguiente grupos: Grupo total (el grupo total de estudiantes), Grupo bajo (estudiantes de 27% con puntuaciones totales más bajas) y Grupo alto (estudiantes de 27% con puntuaciones totales más altas). En "Clave", en la columna de la extrema derecha, un asterisco (*) indica cuál de las alternativas está marcada como la respuesta correcta.

En el reactivo 6, Prop. correcta es .56, es decir, 56% de los estudiantes respondió correctamente este reactivo. Podemos notar que es la misma cifra que aparece en Prop. de aprob. de Alt. 1, es decir, la proporción del grupo total que eligió la respuesta correcta. El Índice de disc. del reactivo 6 es .50; ésta es la diferencia (con cierto error rondando) entre la proporción de aprobación del grupo alto y el bajo en la alternativa 1. Así, en este reactivo, 87% de los estudiantes con mejores puntuaciones totales eligieron la opción correcta, mientras que sólo 36% de los estudiantes con las puntuaciones más bajas eligieron esta opción. Este reactivo fue muy eficaz al separar los grupos alto y bajo. La opción 2 fue atractiva para los estudiantes del grupo bajo, pues casi la mitad de ellos la eligieron; incluso algunos estudiantes (pero muy pocos) del grupo alto eligieron la opción 2 en este reactivo. La correlación biserial puntual (.43) no puede determinarse directamente de los datos que tenemos aquí, pero es la correlación entre la puntuación total de la prueba y el desempeño en este reactivo. La correlación biserial y el índice de discriminación, por lo general, son similares, como se ilustra con los ejemplos del cuadro

6-9.

El reactivo 10 tiene casi el mismo nivel de dificultad que el 6 (.62 vs .56, no hay mucha diferencia), pero el 10 tuvo mucho menor poder de discriminación que el reactivo 6. Mientras que 64% del grupo bajo eligió la opción correcta (Alt. 2), sólo un porcentaje ligeramente mayor (73%) del grupo alto eligió esta opción. Cerca de la tercera parte de cada grupo eligió la opción 4. Estos datos sugieren que el reactivo 10, en especial la opción 4, debe revisarse con mucho detenimiento.

El reactivo 23 es muy difícil; sólo 25% del grupo total lo respondió correctamente. Ningún estudiante del grupo bajo acertó en este reactivo. Aunque éste muestra una discriminación excelente, el hecho de que más estudiantes del grupo alto que del grupo bajo eligieran la opción 4 nos hace preguntarnos sobre esa opción.

El reactivo 28 es muy fácil; casi todos lo respondieron de manera correcta. Es útil como validación de que los estudiantes aprendieron el contenido del reactivo. Sin embargo, contribuye poco a distinguir entre los que saben más o menos del material, como lo indica su bajo índice de discriminación.

El reactivo 29 tiene dificultad moderada (valor p = .69), pero la distribución de las respuestas a lo largo de las opciones es desconcertante. El índice de discriminación y la correlación biserial puntual están cerca del cero. La división entre grupos alto y bajo es más o menos la misma en cada opción. La redacción de este reactivo debe examinarse.

¡Inténtalo!

A continuación aparecen algunos datos de reactivos ordenados de la misma manera que los del cuadro 6-9. Anota las cifras faltantes de las columnas "Prop. correcta" e "Índice de disc."

Reactivo	Estadísticos del reactivo			Estadísticos de las alternativas				
					Prop. de aprobación			
	Prop. correcta	Índice de disc.	Biser. puntual	Alt.	Total	Inferior	Superior	Clave
3	—	—	.48	1	.00	.00	.00	
				2	.15	.36	.13	
				3	.85	.64	.87	*
				4	.00	.00	.00	

Estadísticos del reactivo en la teoría de la respuesta al reactivo

La discusión de los estadísticos del reactivo de la sección anterior se basó en la teoría clásica de las pruebas (TCP). Los índices de dificultad y de discriminación del reactivo en esta teoría a menudo se denominan estadísticos *tradicionales* del reactivo. La teoría de la respuesta al reactivo (TRR) también utiliza estadísticos del reactivo, pero los conceptos y la terminología son algo diferentes a los de la TCP.

Una característica clave del análisis de reactivos en la TRR es la **curva característica del reactivo (CCR)**, la cual relaciona el desempeño en un reactivo con el estatus del

rasgo o capacidad que subyace en la escala. El desempeño en el reactivo se define como la **probabilidad de pasar** un reactivo; *pasar* significa dar la respuesta correcta, si se trata de una prueba de capacidad o aprovechamiento, o responder en cierta dirección, si se trata de una prueba de personalidad, intereses o actitudes. El estatus del rasgo se define en términos de theta ([θ]), como se discutió en el capítulo 3. (La terminología de la TRR se originó primordialmente del trabajo con las pruebas de capacidad y aprovechamiento; a menudo se transfiere de manera directa, forzando un poco el significado, a las medidas de personalidad, intereses y actitudes. Así, theta representa una "capacidad" aunque el constructo sea depresión o interés en la política; el desempeño en el reactivo se califica como "pasar" aunque la respuesta sea "Sí" o "Me gusta". Algunos autores usan los términos más genéricos "rasgo" o "probabilidad de la respuesta correcta".) Los valores theta son hasta cierto punto arbitrarios, pero por lo general varían entre −4 y +4, donde los valores negativos representan menos del rasgo y los positivos, más. La CCR es un esquema de la relación entre estos dos constructos.

La figura 6-6 describe cuatro CCR. Los niveles crecientes de [θ] (es decir, el desplazamiento de izquierda a derecha a lo largo de la base de cada curva) reflejan el aumento en la probabilidad de pasar el reactivo. Esto es cierto para las cuatro CCR. Podemos notar las líneas punteadas asociadas con la CCR del reactivo A; la línea horizontal muestra el punto en que la curva supera 50% de probabilidades de pasar, mientras que la línea vertical muestra dónde se ubica este punto en la escala theta (−1.5 en este ejemplo). Es decir, teóricamente, las personas con −1.5 del rasgo tienen 50% de probabilidades de pasar el reactivo. En el reactivo A, las personas con [θ] = −2.5 y [θ] = 0.0 tienen 20% y 95% de probabilidades de pasarlo, respectivamente. Aquí usamos los valores originales theta, pero en el trabajo práctico, solemos agregar una constante (p. ej., +5) para eliminar los números negativos.

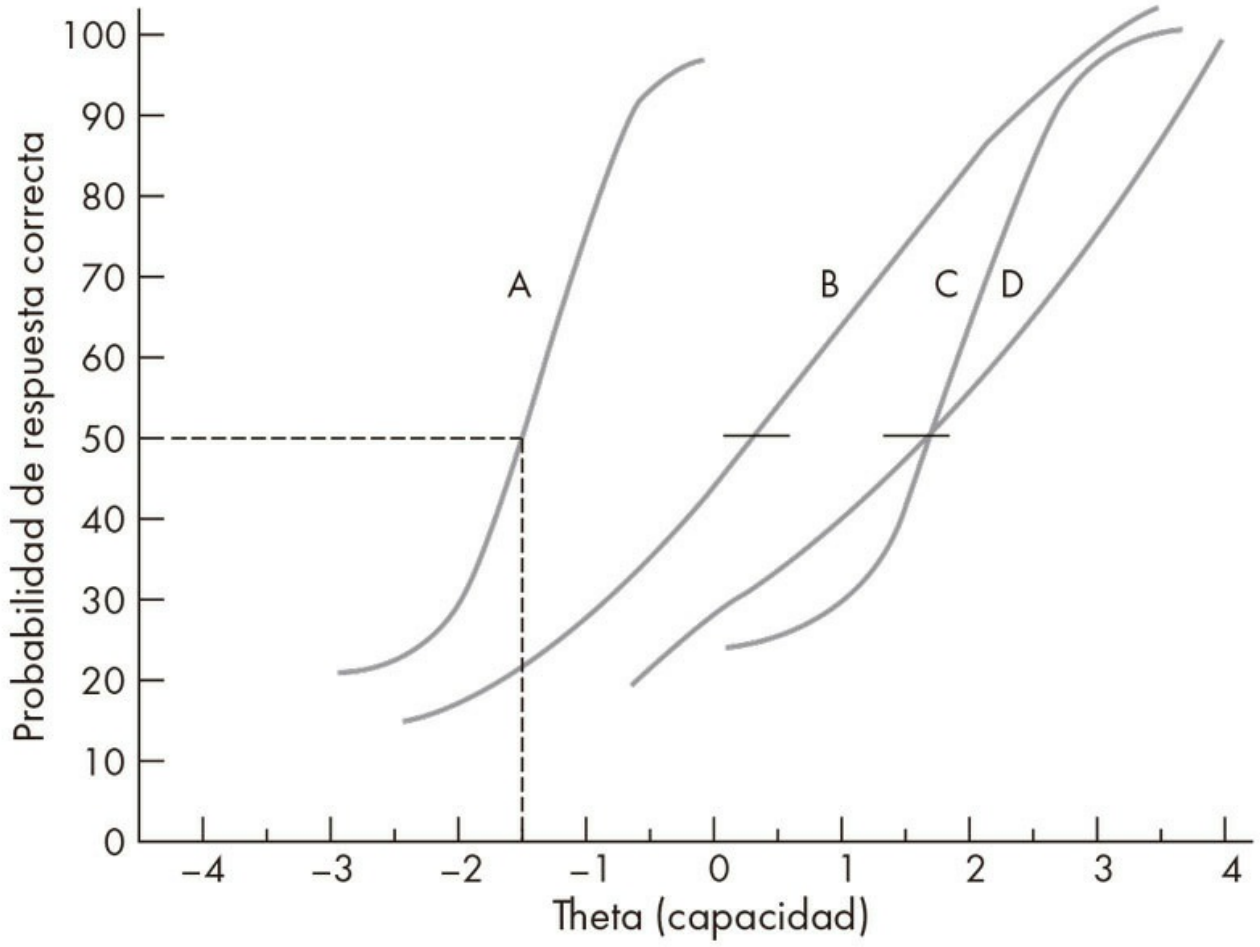


Figura 6-6. Ejemplos de la curva característica del reactivo (CCR) en la teoría de la respuesta al reactivo.

En los modelos más comunes de la TRR, el *parámetro de dificultad* del reactivo es el punto en el que la CCR supera 50% de probabilidades de pasar el reactivo. Este concepto es muy parecido al del índice de dificultad del reactivo (valor p) en la TCP; sin embargo, el parámetro de dificultad en la TRR se representa mediante su valor correspondiente [θ].

Podemos observar las marcas en las CCR de los reactivos B, C y D en la figura 6-6, las cuales muestran el punto en el que las curvas superan el 50% en el eje *y*. El reactivo B es más difícil que el A; se necesita un nivel de capacidad superior (θ) para tener 50% de probabilidades de pasar B en comparación con A. Los reactivos C y D superan la marca de 50% en el mismo punto, y ambos son más difíciles que A y B.

Las CCR de la figura 6-6 no tienen todas la misma forma; las de los reactivos A y C son muy parecidas: tienen una marcada forma de "S". La mitad de la curva, en ambos casos, es muy empinada. Las CCR de los reactivos B y D son menos pronunciadas; el término formal para esta inclinación es **pendiente**. La pendiente de una curva muestra qué tan marcadamente el reactivo diferencia entre las personas con capacidades distintas

(valores θ). Hablar de "diferenciar" nos recuerda al índice de discriminación de la TCP; de hecho, la pendiente de la CCR corresponde de manera estrecha a esta noción.

El reactivo A muestra una marcada diferenciación de −2.0 a −.5 en la escala [θ]. El reactivo C tiene la misma pendiente que el reactivo A, pero el C funciona mejor (es decir, discrimina de modo más claro) en el rango de 1.0 a 3.0. Esto ilustra una característica importante de la CCR: ayuda al creador de pruebas a identificar los reactivos que funcionan de manera diferenciada en distintos puntos del espectro de la capacidad.

Hay una tercera característica de la CCR en la figura 6-6. El extremo inferior de la curva de los reactivos A y C se hace más plano alrededor del nivel de 20% del eje *y*. Desde el punto de vista técnico, este "aplanamiento" se conoce como asíntota inferior. Así, sin importar qué tan abajo se encuentre una persona en [θ], hay cerca de 20% de probabilidades de pasar el reactivo. Al principio, esto puede parecer inexplicable; sin embargo, consideremos el caso de un reactivo de opción múltiple con cinco opciones de respuesta. Sin importar cuán poco sepamos acerca del tema que se está midiendo, tenemos 20% de probabilidades de pasar el reactivo tratando de adivinar al azar. Algunos modelos de la TRR explican esta asíntota inferior con el **parámetro de adivinación**.[7] En el caso de un reactivo de 10 opciones de respuesta, el parámetro de adivinación es posible que esté al nivel de 10%. Podemos notar que el extremo inferior de la CCR del reactivo B se aproxima a cero en el eje *y*, lo cual ilustra que la adivinación no afecta todos los reactivos. Por lo general, las asíntotas superiores de la CCR están cerca de 100%; de ahí que no se introduzca un parámetro separado en el caso de éstas.

Hemos examinado tres parámetros de una CCR: dificultad, pendiente y adivinación. En el lenguaje de la TRR, la pendiente o parámetro de discriminación es "a", el parámetro de dificultad es "b" y el parámetro de adivinación es "c". Estos parámetros dan origen a tres modelos de la TRR, a menudo denominados P1, P2 y P3: modelos de uno, dos y tres parámetros. El modelo de un parámetro toma en cuenta sólo el parámetro de dificultad (b); este modelo supone que todos los reactivos tienen la misma pendiente (poder de discriminación) y que la adivinación no es un factor significativo. El modelo más popular de un parámetro es el **modelo Rasch**, nombrado así por George Rasch, quien lo desarrolló (Wright, 1997). El modelo de dos parámetros toma en cuenta dificultad y discriminación, pero no la adivinación, mientras que el modelo de tres parámetros toma en cuenta dificultad, discriminación y adivinación. (Véase los ejercicios 13 y 14 al final del capítulo para experimentar con los tres parámetros.)

La figura 6-7 muestra las CCR de dos reactivos de un proyecto real de elaboración de una prueba que usa el modelo Rasch. Los puntos conectados muestran el desempeño real de los subgrupos en el programa de investigación. Las CCR se ajustan a estos puntos empíricos. El reactivo 40 es relativamente fácil, pues su parámetro de dificultad (b) es de −2.70. El reactivo 352 es más difícil: b = 1.67.

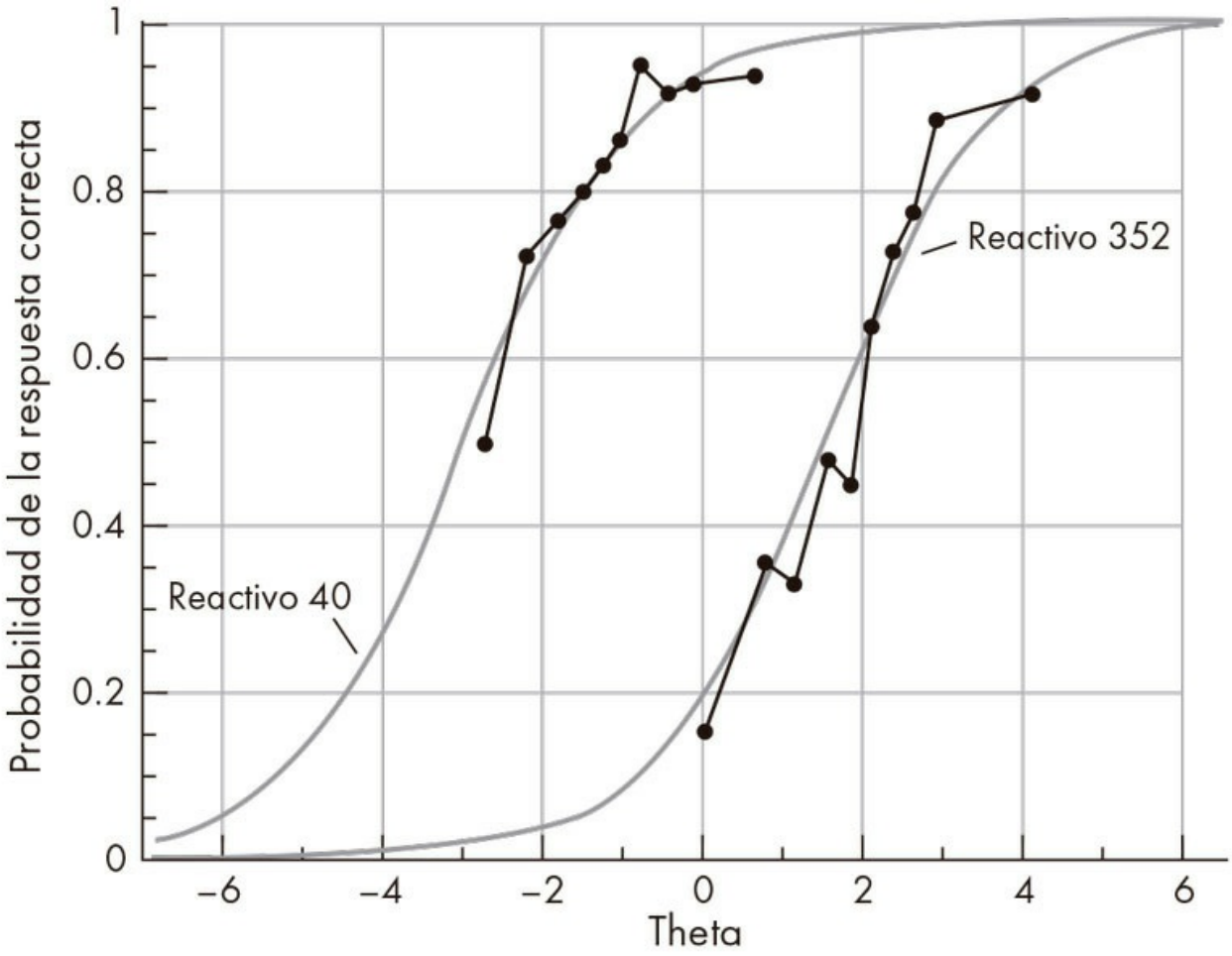


Figura 6-7. Ejemplos de la CCR para reactivos que usan el modelo Rasch.
Fuente: Reproducida con autorización de Renaissance Learning, Inc.

¡Inténtalo!

Traza una línea recta en las CCR de la figura 6-7; verifica que los valores theta sean −2.70 y 1.67 en la "probabilidad de respuesta correcta" de .50.

Para comprender las CCR, puede ser útil introducir algunos ejemplos que son teóricamente posibles aunque poco probables de ocurrir en la práctica. La figura 6-8 muestra estos ejemplos. El reactivo E muestra un caso en el que todos los que están debajo de cierto nivel de capacidad ([θ] = −2.0) fallan y todos los que están arriba de ese nivel lo pasan. Desde varios puntos de vista, éste es un reactivo ideal. Una serie de reactivos como éste en diferentes niveles de [θ] harían una prueba muy eficiente. El reactivo F muestra un caso en el que el reactivo diferencia positivamente hasta cierto punto y, luego, pierde su poder de diferenciación; después, lo recupera de nuevo. A veces encontramos un patrón como éste, pero es, quizá, más una cuestión de

fluctuaciones inestables en la muestra que un verdadero fenómeno. El reactivo G muestra el extraño caso en que la probabilidad de pasar el reactivo en realidad disminuye conforme el nivel de capacidad aumenta. Esto correspondería a un índice de discriminación *negativo* en la TCP: más personas del grupo inferior que del grupo superior pasan este reactivo. En realidad, la gráfica del reactivo G no es tan rara como podría parecer a primera vista. En situaciones prácticas es justo la clase de gráfica que ocurre con las opciones *incorrectas* de un reactivo de respuesta cerrada; es decir, conforme aumenta el nivel de capacidad, la probabilidad de elegir una opción incorrecta disminuye.

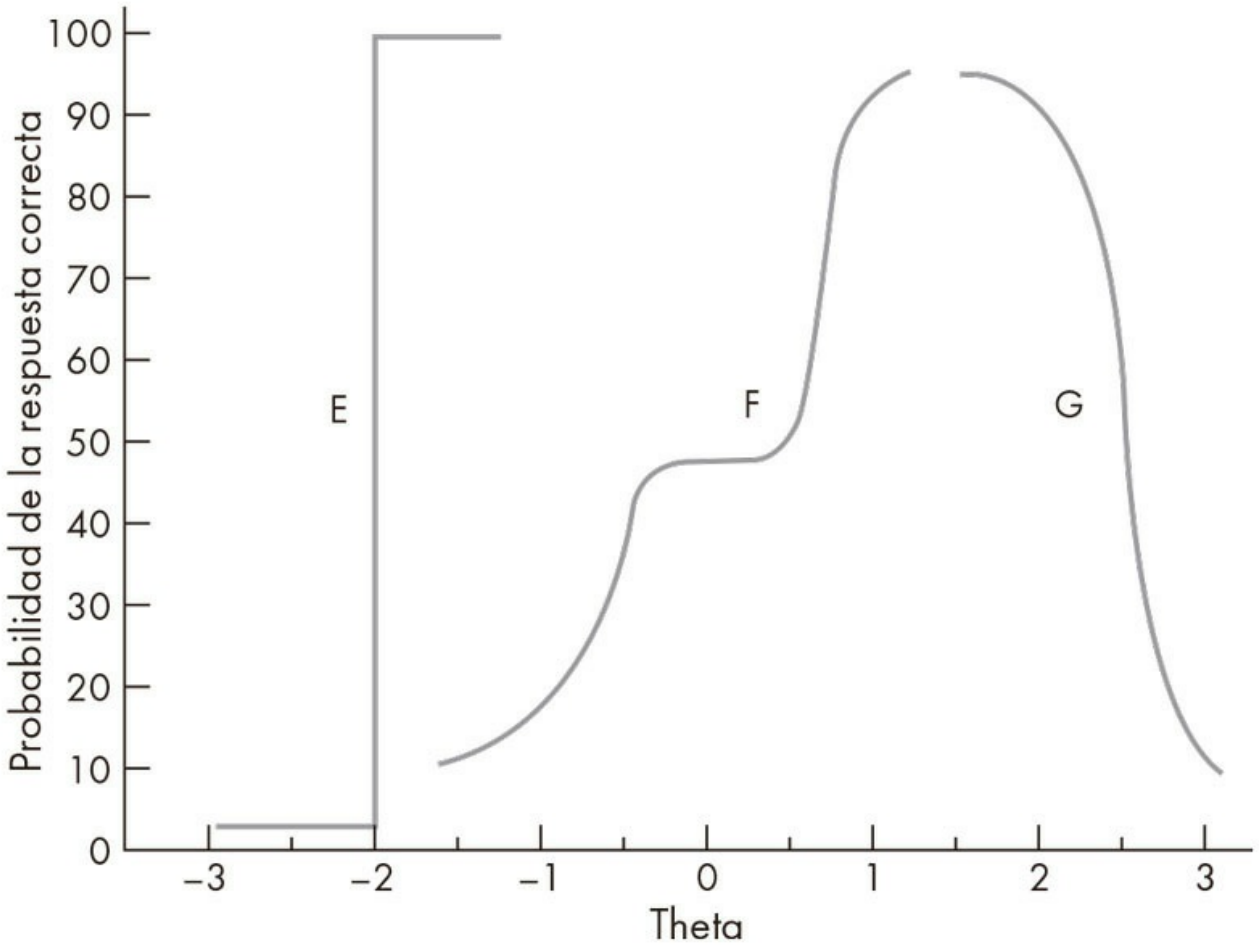


Figura 6-8. Algunas CCR teóricamente posibles, pero poco probables.

Los parámetros de una CCR pueden traducirse como lo que se denomina **función informativa del reactivo**, la cual muestra en qué parte del continuo del rasgo ([θ]) un reactivo proporciona información pertinente para la medición. La figura 6-9 presenta las funciones informativas hipotéticas de dos reactivos. La función del reactivo B muestra que proporciona una cantidad moderada de información a lo largo del rango 0.0–3.0; la cantidad de información es más o menos la misma en la mitad de este rango. El reactivo

A está marcadamente enfocado alrededor de [θ] = −1.0. La cantidad de información que proporciona A decae con rapidez más allá de −1.0.

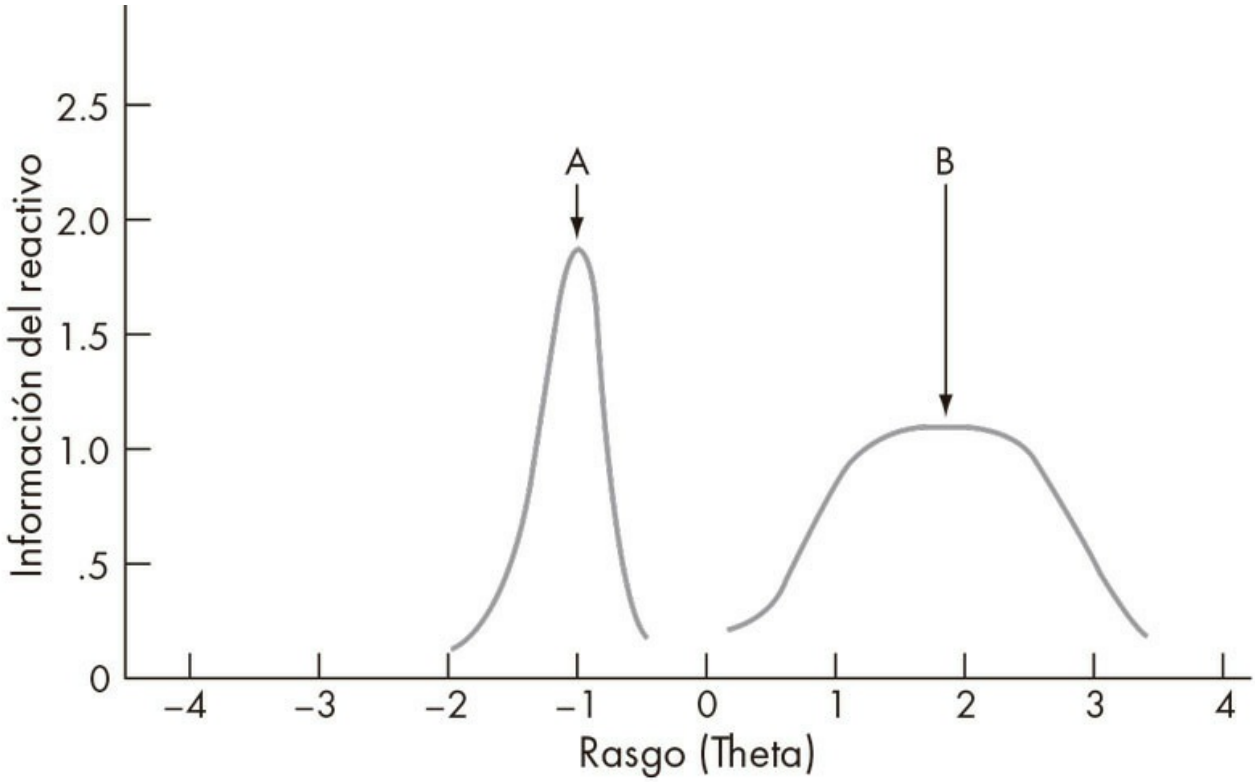


Figura 6-9. Funciones informativas de dos reactivos hipotéticos.

¡Inténtalo!

Con base en lo que sabes sobre las CCR, ¿puedes bosquejar una para los reactivos cuya información se presenta en la figura 6-9?

El valor relativo de los estadísticos del reactivo en la TRR y la TCP es un poco polémico. La mayoría de los creadores contemporáneos de pruebas emplean los estadísticos de la TRR, lo que sugiere que los expertos encuentran valiosos los datos que se obtienen de ese modo. Sin embargo, los estadísticos de la TCP siguen usándose en los mismos proyectos de elaboración de pruebas. Esto parece ser más que sólo una cuestión de aferrarse a lo conocido por la seguridad que brinda. Los creadores de prueba activos encuentran valor en los estadísticos tradicionales, así como en los de la TRR.

Análisis factorial como técnica de análisis de reactivos

En el capítulo 5, describimos el análisis factorial como un método que se usa para

demostrar la validez de constructo de una prueba. En esa aplicación, la prueba ya estaba hecha y el análisis factorial indicaba su estructura. Sin embargo, esta técnica también se usa en la fase de análisis de reactivos para ayudar a elegir los reactivos que producirán puntuaciones relativamente independientes y significativas. Este enfoque se usa mucho en la construcción de escalas multirrasgo de personalidad, intereses y actitudes.

En esta aplicación del análisis factorial, una gran cantidad de reactivos claramente pertinentes a los rasgos que se tiene pensado medir se aplica a una muestra de examinados. Las intercorrelaciones entre los reactivos se analizan factorialmente. Se identifican las dimensiones subyacentes (factores). Entonces, los reactivos con cargas altas en los factores se eligen para formar parte de la prueba final, la cual produce puntuaciones separadas de cada uno de los factores con reactivos que deben ser medidas relativamente puras y eficientes de los factores.

Cuadro 6-10. Resultados parciales de un análisis factorial de reactivos de un inventario de intereses

Reactivo	**Factor: I**	**II**	**III**	**IV**
			Cargas de los reactivos[a]	
1	10	**76**	−07	06
2	05	16	10	**73**
3	08	29	**59**	**39**
4	19	**39**	**67**	−05
5	**51**	26	**47**	−11
6	**36**	**51**	**33**	**−31**
7	12	**44**	**40**	17
8	03	24	**65**	−01
9	09	06	**55**	16
10	**58**	**45**	23	01

[a] Se omitieron los puntos decimales.

El cuadro 6-10 muestra parte de los resultados del análisis factorial de un fondo de reactivos diseñados para medir los intereses de los niños en varios temas escolares. Los reactivos con cargas en negritas podrían elegirse para las escalas finales. En este proceso, las cargas de los reactivos en los factores sirven para un propósito similar al de los índices de discriminación (D); sin embargo, para determinar D, necesitamos una puntuación total para dividir la muestra de prueba en grupos alto y bajo. En el método del análisis factorial, no tenemos puntuaciones totales para empezar el proceso, sino que generamos factores y determinamos la relación entre los reactivos y dichos factores.

Selección de reactivos

La fase final del proceso de análisis de reactivos es la selección. De todos los reactivos preparados y probados, se seleccionan los que aparecerán en la prueba para su

estandarización.[8] La selección de reactivos toma en cuenta el propósito y diseño originales de la prueba, las especificaciones pertinentes de contenido y los datos del análisis de reactivos. Aquí identificamos varias directrices para este proceso, las cuales derivan de los principios que desarrollamos en los capítulos sobre normas, confiabilidad y validez. La selección de reactivos no ocurre en el vacío, sino que las características de una buena prueba regulan este proceso.

1. A menudo, el número total de reactivos de la prueba es lo más importante para determinar su confiabilidad. Desde luego, a todos les gustan las pruebas cortas, pero éstas, por lo general, no son muy confiables. Como regla general, para aumentar la confiabilidad de una prueba, es necesario aumentar el número de reactivos; sin embargo, hay un punto en el que agregar nuevos reactivos no aumenta la confiabilidad de manera significativa.

Al considerar el número deseado de reactivos, la atención se debe concentrar en la puntuación, o puntuaciones, que serán informadas, y no el simple número de reactivos en la prueba. Supongamos que la prueba tiene 100 reactivos, pero las puntuaciones importantes se basan en seis grupos de ellos, uno de los cuales tiene 50 y los otros cinco tienen 10 reactivos. El grupo de 50 reactivos es, probablemente, el que proporcione una puntuación confiable, mientras que los otros cinco grupos probablemente proporcionen puntuaciones no confiables. Así, el hecho de que la prueba tenga 100 reactivos es, por completo, irrelevante.

2. El nivel de dificultad promedio de la prueba está en función directa de los valores p de los reactivos, mientras que la puntuación media de la prueba es sólo la suma de los valores p. Otro modo de expresar esto es que la puntuación media de la prueba es el promedio de la multiplicación de los valores p por el número de reactivos de la prueba. Como señalamos antes, el valor p es, en realidad, un índice de la facilidad del reactivo más que de su dificultad. De ahí que para obtener una prueba fácil se usen reactivos con valores p altos, mientras que parar obtener una prueba difícil se usen reactivos con valores p bajos. Depende del propósito de la prueba si se desea que sea fácil o difícil; una prueba fácil ofrecerá la mejor discriminación en el extremo inferior de la distribución de las puntuaciones de la prueba, mientras que una prueba difícil lo hará en el extremo superior. Se puede desear una prueba fácil para una prueba diagnóstica de lectura diseñada para dar buena información sobre estudiantes con dificultades para leer. La figura 6-10, Prueba A, ilustra la distribución de puntuaciones de una prueba así. En esta distribución, el rango entre los casos es más grande en la porción inferior; este tipo de distribución es resultado de tener muchos reactivos con valores p altos. Por otro lado, puede desearse una prueba que despliegue los casos en la parte superior de la distribución, por ejemplo, para elegir candidatos a una beca. La distribución deseada para este caso se muestra en la figura 6-10, Prueba B. Este tipo de distribución es resultado de tener muchos reactivos con valores p bajos. En la terminología estadística, la distribución de la prueba A tiene una asimetría negativa o hacia la izquierda, mientras que la de la prueba B tiene una asimetría positiva o hacia la derecha. Debe quedarnos claro que no hay una regla de que las pruebas psicológicas

inevitablemente lleven a una distribución normal de las puntuaciones.

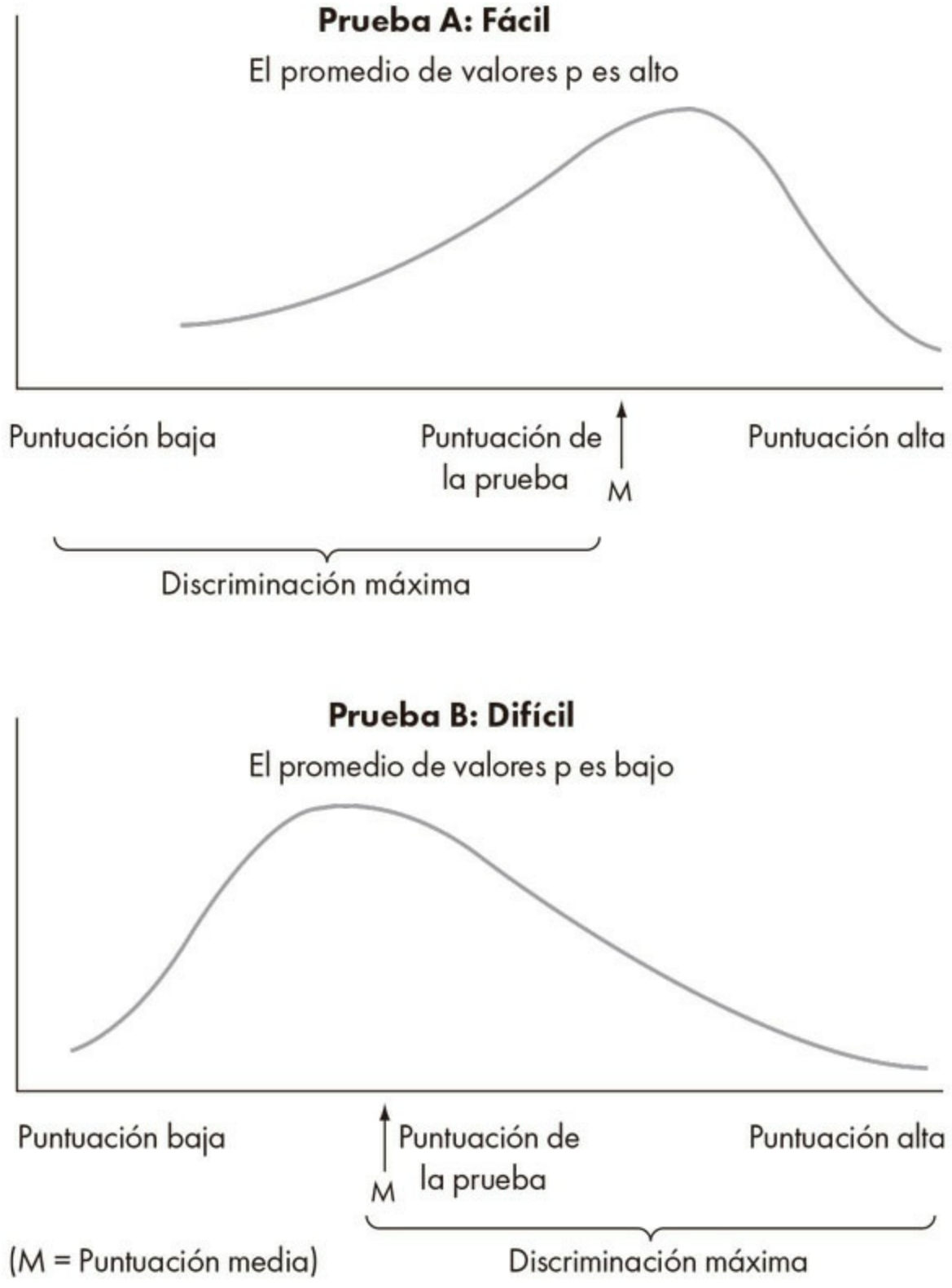


Figura 6-10. Distribuciones que resultan de elegir reactivos con valores p altos o bajos.

3. En general, queremos reactivos con índices de discriminación altos.[9] Tales reactivos contribuyen a la medición del rasgo. ¿Qué es "alto" para un índice de discriminación? Mientras que solemos pensar en correlaciones "altas" de .80 o más, un buen (alto) índice de discriminación a menudo no es mayor de .50, y un índice de .30 es bastante respetable. Necesitamos recordar que un solo reactivo tiene una confiabilidad muy limitada, por lo que es poco probable tener una correlación sumamente alta con cualquier otra variable. Sin embargo, un conjunto de muchos reactivos con índices de discriminación de .30 a .50 conformará una prueba muy buena. Esta directriz se aplica sin importar el método con que se determine la discriminación del reactivo. Sin duda, los índices de discriminación negativa deben evitarse. Los reactivos con índices de discriminación cercanos a cero no contribuyen en nada a la medición del rasgo.

Cuando se usa el análisis factorial como técnica de análisis de reactivos, la meta suele ser obtener varias pruebas correspondientes a los factores, que son medidas relativamente "puras" de los factores. Por lo tanto, elegimos reactivos que tengan cargas altas en un solo factor y cargas relativamente bajas en todos los demás factores.

4. Hay una relación importante entre el valor p de un reactivo y el índice de máxima discriminación posible (D). De manera específica, D puede tomar su valor máximo cuando p está en su punto medio. Consideremos los ejemplos del cuadro 6-11 de un grupo total de 200 casos. Con estos datos, D se basa en una división del grupo total en 50% superior e inferior, lo que nos da 100 casos para cada grupo. El cuadro muestra el número de casos de cada grupo que tuvo el reactivo correcto (No. correcto), luego se traduce este número a una proporción (Prop.) de cada grupo. Recordemos que el índice de discriminación es la diferencia entre la proporción de correctos en el grupo superior y el inferior. Si p = 1.00, es decir, todos respondieron correctamente al reactivo, es decir, 100% del grupo superior y del inferior, entonces D = .00. Un análisis muy similar se puede hacer si todos respondieron incorrectamente: el valor p es .00 y D = .00. Si el valor p es .50, la proporción de casos del grupo alto puede ser de 1.00 y la del grupo bajo, de .00; de este modo se obtiene el valor máximo de D = 1.00. Si p = .50, se puede obtener la diferencia máxima entre los grupos superior e inferior. Examinemos otras combinaciones en el cuadro 6-11 para confirmar la relación entre p y D.

Cuadro 6-11. Ejemplos de la relación entre el valor p de un reactivo y el índice de discriminación máxima posible

Grupo total (N = 200)		**Grupo inferior (N = 100)**		**Grupo superior (N = 100)**		
No. correcto	**Prop. (p)**	**No. correcto**	**Prop.**	**No. correcto**	**Prop.**	**Máxima posible D**
200	1.00	100	1.00	100	1.00	.00
150	.75	50	.50	100	1.00	.50
125	.625	25	.25	100	1.00	.75

100	.50	0	.00	100	1.00	1.00
60	.30	0	.00	60	.60	.60
40	___	___	___	___	___	___
0	.00	0	.00	0	.00	.00

¡Inténtalo!

Para asegurarnos de que comprendes el cuadro 6-11, anota los valores faltantes de 40.

Debemos señalar dos cuestiones acerca de la relación entre p y D. Primero, esta relación se refiere a la máxima *posible* D. Sin duda, podría darse el caso de que p = .50 y 50% de los casos de los grupos alto y bajo respondieran correctamente, lo que produciría que D = .00. El valor p determina qué tan alta *puede* ser D, no qué tan alta en realidad es; sin embargo, por lo general, sucede que en la práctica, al elaborar una prueba, hay una relación real entre p y D.

Segundo, cuando fijamos el punto medio del valor de p = .50, suponemos que no hay efecto de adivinación, pero en la práctica, por lo común, está presente en cierto grado en las pruebas de opción múltiple. Con propósitos semejantes al del análisis incluido en el cuadro 6-11, el punto medio se define como la marca intermedia entre una puntuación perfecta y una puntuación por azar. Por ejemplo, en el caso de una prueba con reactivos de cinco opciones, donde la puntuación por azar es de 20%, el punto medio del valor de p es .60, no .50.[10]

Habiendo considerado la relación entre p y D, regresemos al tema de la distribución de los valores p deseados de la prueba. Ya que D puede maximizarse cuando p = .50 (o ajustados adecuadamente hacia arriba por la adivinación), a veces la recomendación es que se elijan para la prueba los reactivos con p = .50. Ésta ha sido una recomendación influyente en el campo de las pruebas, pues ayuda a explicar por qué las pruebas de capacidad y aprovechamiento parecen tan difíciles a los examinados. Tener incorrecta la mitad de reactivos de una prueba es una experiencia perturbadora.

No todos los expertos concuerdan con la recomendación de elegir la mayoría de reactivos con valores p = .50. Esta recomendación sólo se puede aplicar cuando se desea hacer la máxima discriminación en la mitad de la distribución de las puntuaciones. Esto a veces es deseable, pero a veces no. En muchas situaciones de evaluación, deseamos hacer discriminaciones razonablemente buenas a lo largo de todo el rango del rasgo que medimos. Esto sugiere dispersar los valores p de abajo hacia arriba; con esta estrategia, lo que tratamos de hacer, en esencia, es obtener discriminaciones razonablemente buenas en varios puntos al mismo tiempo que sacrificamos la discriminación máxima en un punto. Este análisis otra vez ilustra la influencia que el propósito de la prueba tiene en la manera en que ésta se elabora.

5. Los criterios estadísticos deben atemperarse siguiendo consideraciones no estadísticas en la selección de reactivos. En una prueba de aprovechamiento, pueden incluirse ciertos reactivos para satisfacer las demandas de las especificaciones del

contenido de la prueba, es decir, para asegurar la validez de contenido. Por ejemplo, el proyecto de una prueba de matemáticas puede exigir 10 reactivos sobre conceptos y 10 sobre cálculos. Incluso si los estadísticos son más favorables en el caso de los reactivos de conceptos que los de cálculo, no excluiríamos estos últimos reactivos.

También se podrían incluir reactivos con fines motivacionales. Por ejemplo, es común empezar las pruebas de capacidad con reactivos muy fáciles para que el examinado pueda tener un buen inicio. Los reactivos pueden tener valores p de .99 y, por lo tanto, los valores de D estarán cerca de .00. Sin embargo, los reactivos aún tienen un propósito útil en la prueba; en los inventarios de personalidad e intereses, los reactivos a veces se repiten deliberadamente para verificar la consistencia en las respuestas.

Programas de estandarización e investigación complementaria

El **programa de estandarización** produce las normas de una prueba; a veces se le llama así y a veces también se le llama programa de obtención de normas. Este programa es parte importante de la elaboración completa de una prueba; se realiza después de que se seleccionaron los reactivos en la fase final de la etapa de análisis de reactivos. La prueba que se estandariza debe ser la prueba exacta que se publicará al final. Todas las instrucciones, el número de reactivos, los límites de tiempo, deben estar determinados; de lo contrario, las normas que resulten correrán peligro.

En el capítulo 3, describimos la naturaleza de los programas de estandarización en relación con el tratamiento de las normas. No es necesario repetir esa presentación aquí, por lo que nada más señalamos el lugar que ocupa la estandarización en el proceso de elaboración de una prueba.

Sea como parte del programa de estandarización o como algo simultáneo a ésta, por lo general, hay otros programas de investigación realizados con la versión final de la prueba antes de su publicación. La naturaleza y extensión de estos programas dependen del alcance de la prueba; aquí sólo mencionaremos algunos de los programas que pueden llevarse a cabo.

Algunos programas de investigación se realizarán sólo analizando los datos del programa de estandarización. Estos programas son independientes en términos lógicos de la elaboración de normas –que es el principal propósito del programa de estandarización–, pero no requieren que se levanten datos nuevos. A menudo, los análisis de puntuaciones de acuerdo con género, raza, edad, región geográfica y otras variables demográficas se hacen con los datos de estandarización; también se pueden hacer estudios de la validez de la prueba. Las relaciones de la prueba con otras o con valoraciones de supervisores, clínicos o maestros pueden obtenerse para las submuestras del grupo de estandarización. La estructura analítico-factorial de la prueba puede determinarse con los datos de estandarización.

Ahora se pueden llevar a cabo varios tipos de estudios de confiabilidad, por ejemplo, de test-retest. Por lo general, no es factible hacer un estudio como éste con todos los que participaron en el programa de estandarización; sin embargo, se puede aplicar otra vez la prueba a una submuestra en una fecha posterior. Ya que los estudios de test-retest son onerosos, pueden realizarse con una muestra por completo independiente del grupo de estandarización. Si la prueba tiene más de una forma, podría llevarse a cabo un estudio de confiabilidad de formas alternas junto con el programa de estandarización o un estudio independiente paralelo a dicho programa. Las medidas de consistencia interna, por ejemplo, el coeficiente alpha, se hacen con facilidad en la muestra de estandarización completa; se trata sólo de un análisis estadístico que no requiere recopilar nuevos datos.

Se pueden llevar a cabo tres tipos de *programas de igualación* como parte o, al

menos, al mismo tiempo que el de estandarización: igualación de formas alternas de la prueba (si están disponibles), igualación de niveles diferentes de la prueba (si la prueba tiene niveles múltiples) e igualación de ediciones nuevas y antiguas (si la prueba es una revisión). Kolen y Brennan (2004) son una referencia clave de los programas de igualación.

Preparación de los materiales finales y publicación

El paso final en el proceso de elaboración de una prueba es su publicación. ¿Qué es exactamente lo que se publica? En el uso cotidiano de la palabra *"publicado"*, tendemos a pensar en la impresión del cuadernillo de la prueba o el conjunto de estímulos, como las láminas del TAT. Pero la publicación de una prueba implica instrucciones de aplicación e interpretación, manuales técnicos, informes de puntuaciones y otros materiales complementarios. En las pruebas en verdad sencillas de uso limitado, el conjunto de los materiales puede ser bastante modesto: un cuadernillo, una clave de calificación y un manual de 20 páginas con las instrucciones de aplicación y las características técnicas de la prueba. En el caso de pruebas complejas que se usan mucho, el conjunto de los materiales puede ser asombrosamente grande; puede incluir varios tipos de manuales, materiales interpretativos complementarios, informes técnicos especiales, programas de cómputo complejos para calificar y hacer informes, y versiones de la prueba en lenguas extranjeras y ediciones en Braille.

Una prueba publicada debe tener un *manual técnico*, el cual es la fuente clave de información acerca del propósito, fundamentos y estructura de la prueba. El manual debe incluir información sobre la confiabilidad, validez y procedimiento de estandarización de la prueba. Por último, también debe incluir directrices para interpretar la puntuación o puntuaciones. Algunas pruebas pueden cubrir todos estos rubros en un manual, mientras que otras pueden tener más de uno.

Muchas pruebas actuales tienen **informes de puntuaciones**, los cuales pueden incluir la presentación gráfica de las puntuaciones y/o la traducción de las puntuaciones numéricas a una forma narrativa. Las pruebas más usadas de aplicación grupal de aprovechamiento y capacidad producen informes generados por computadora no sólo sobre un individuo, sino también sobre un grupo, por ejemplo, de un salón de clases, de una escuela, del sistema escolar entero o incluso de todo un estado.

Por último, la publicación puede suponer diversos **materiales complementarios.** Por ejemplo, algunas pruebas tienen "localizadores" que ayudan al aplicador a determinar cuál es el nivel más apropiado para un examinado cuando la prueba es de múltiples niveles. Algunas pruebas ofrecen cuadernillos especiales sobre la interpretación de las puntuaciones para los estudiantes y sus padres.

En realidad, puede ser un poco engañoso identificar la publicación con el último paso del proceso de elaboración de una prueba, pues éste nunca termina. Cuando una prueba se publica, nunca cuenta con una demostración exhaustiva de su validez; además, siempre hay más preguntas acerca de su aplicabilidad en varias poblaciones especiales. Sin importar qué tan perfecto haya sido el programa de estandarización, las normas están ligadas al tiempo, por lo que hay una preocupación constante relacionada con lo posibilidad de que éstas se vuelvan anticuadas debido a cambios en la población meta.

Por todas estas razones, una prueba estará sujeta a investigación adicional aun después de su publicación. Parte de este desarrollo estará a cargo del autor o autores y la editorial, pero también otros usuarios interesados emprenderán estudios sobre la prueba; algunos estudios serán publicados en revistas dedicadas a las pruebas como mencionamos en el capítulo 2.

> Las pruebas y los documentos que las sustentan... se revisan de manera periódica para determinar si se requiere una revisión. Las revisiones o enmiendas son necesarias cuando nuevos datos de investigaciones, cambios significativos en el dominio o nuevas condiciones del uso e interpretación de la prueba mejorarían la validez de las puntuaciones o sugieren que la prueba ya no está en condiciones óptimas o no es por completo apropiada para los usos para los que fue pensada.
>
> *Standards...* (AERA, APA, & NCME, 2013)

Neutralidad y sesgos «160-169a

Ahora retomaremos el tema de la neutralidad de la prueba. Como señalamos antes, el tema de la neutralidad pertenece, en términos lógicos, al capítulo 5 sobre validez, y veremos por qué es así en un momento. Sin embargo, retrasamos la revisión de este tema hasta el final de este capítulo por una razón muy práctica. Muchos de los esfuerzos para asegurar la neutralidad ocurren durante la elaboración de la prueba; por ello, teníamos que conocer el procedimiento normal que se sigue para crear una prueba con el fin de comprender algunos de los procedimientos para tratar con la neutralidad.

La neutralidad de la prueba es, con seguridad, uno de los temas más polémicos en el campo de las pruebas psicológicas y educativas; hay mucha confusión alrededor de este concepto y, a menudo, también mucha pasión. De ahí que sea importante empezar nuestro análisis del tema poniéndolo en perspectiva y considerando algunos ejemplos. En la literatura profesional de las pruebas psicológicas, los términos **neutralidad de la prueba** y **sesgo de la prueba**, por lo general, tienen el mismo significado, pero connotaciones opuestas. Una prueba neutral es la que carece de sesgos, mientras que una prueba sesgada es la que carece de neutralidad. Usaremos ambos términos –neutralidad y sesgo– en nuestra discusión.

El tema de la neutralidad en perspectiva

Neutralidad significa que una prueba (o alguna otra técnica de evaluación) mide un rasgo, constructo u objetivo con una validez equivalente en distintos grupos. Una prueba está sesgada (no neutral) si no mide el rasgo de interés de la misma manera en diferentes grupos. Una simple diferencia en el desempeño promedio entre los grupos no constituye un sesgo; éste sólo existe si la diferencia en los promedios no corresponde a una diferencia real en el rasgo subyacente que la prueba trata de medir. Los promedios grupales pueden diferir –de hecho, deben diferir– si los grupos en verdad son diferentes respecto de la capacidad o rasgo que se intenta medir. Para ilustrar este punto importante, examinemos algunos ejemplos. Consideremos el contraste entre alumnos que estudian y otros que no lo hacen para el examen final de un curso sobre pruebas psicológicas. Las personas del grupo A estudian el libro de texto 20 horas a la semana y asisten a todas las clases, mientras que las del grupo B estudian el libro de texto 20 minutos la noche anterior al examen y asisten a clases de manera irregular. En el examen final, el promedio del grupo A es notablemente superior que el del grupo B, pero esa diferencia no significa que el examen tenga un sesgo en contra del grupo B. De hecho, nos sorprendería no encontrar esta diferencia entre los promedios grupales. ¿Por qué? Porque suponemos que hay una diferencia real entre estos grupos respecto del rasgo subyacente de conocimiento sobre el tema. Además, si el desempeño en el examen final se toma como algo que pronostica el desempeño en el GRE *Subject Exam in Psychology,* sin duda predecirá puntuaciones superiores para las personas del grupo A en

comparación con las del grupo B. Esto tampoco indica que el examen final esté sesgado. ¿Por qué? Porque suponemos que las personas que han estudiado más tendrán un mejor resultado en el examen GRE que las que estudiaron menos.

> ... la perspectiva sobre la medición del *Standards excluye* de manera explícita una visión común de la neutralidad en el discurso público: la neutralidad como igualdad de los resultados de las pruebas para subgrupos definidos por raza, origen étnico, género, discapacidad u otras características. Sin duda, la mayoría de los profesionales de la evaluación están de acuerdo en que las diferencias grupales en los resultados de las pruebas deben desencadenar un examen detallado para detectar posibles fuentes de sesgo... Sin embargo, las diferencias grupales en los resultados no indican por sí mismas que una prueba tiene sesgos o no es neutral.
>
> *Standards*... (AERA, APA, & NCME, 2013)

He aquí un segundo ejemplo. Queremos saber si una prueba de lectura es neutral (o está sesgada) en relación con estudiantes con debilidad visual (DV). Comparamos el desempeño de estudiantes con DV con el de estudiantes sin DV; los primeros tienen puntuaciones menores. ¿Eso significa que la prueba no es neutral para los estudiantes con DV? Aún no lo sabemos; puede ser que estos estudiantes en verdad tengan habilidades de lectura inferiores a las del otro grupo. Supongamos que presentamos la prueba en una versión con la letra más grande y encontramos que la puntuación promedio de los estudiantes con DV está por encima de los estudiantes sin DV; esto sugiere que la prueba de lectura, en su versión original con letras pequeñas, no era neutral, sino que estaba sesgada en contra de los estudiantes con DV. El resultado también sugiere que la prueba es neutral cuando se presenta en una versión con letra grande. Pongamos este ejemplo en el contexto de la figura 5-1; el tamaño de la tipografía introduce varianza irrelevante para el constructo, es decir, el tamaño de la letra influye en las puntuaciones, pero no queremos que eso ocurra.

¡Inténtalo!

Las pruebas típicas de comprensión de lectura constan de pasajes seguidos de varias preguntas. La elección de los temas para los pasajes podría sesgar la prueba en favor de los niños o las niñas, digamos, de 12 a 16 años. Menciona tres temas que podrían sesgar la prueba en favor de los niños y otros tres que la podrían sesgar en favor de las niñas.

En favor de los niños	En favor de las niñas
_____	_____
_____	_____
_____	_____

Ahora consideremos el caso de una prueba de sistemas de transporte en EUA. Se espera que los estudiantes sepan acerca de los sistemas de transporte público, así como

de otros sistemas. Los niños rurales pueden tener menos conocimiento sobre el transporte porque, por lo general, no lo usan en su medio. Por lo tanto, los niños rurales tienen puntuaciones menores en los reactivos relacionados con este tema. ¿La prueba no es neutral con los niños rurales? No. Pero lo que queríamos medir era justo el conocimiento sobre los sistemas de transporte público. La solución a este problema es enseñar a los niños rurales sobre los sistemas de transporte público, pues no queremos cambiar la prueba. Por otro lado, supongamos que los reactivos acerca de los sistemas de transporte público tenían la intención de medir la capacidad de lectura. Si los niños rurales tienen puntuaciones menores en la prueba, no a causa de la deficiencia en las habilidades de lectura sino por el desconocimiento del tema, entonces podríamos estar dispuestos a cambiar los reactivos.

La figura 6-11 sugiere cómo pensar sobre estas situaciones. Comparamos los grupos A y B. En la parte alta de la figura, los grupos A y B difieren en el desempeño en la prueba: el grupo A es superior. Sin embargo, los grupos también difieren en términos del estatus real del rasgo. Por lo tanto, la prueba es neutral, no tiene sesgos; sólo refleja las diferencias reales entre los grupos. En la parte inferior de la figura, los grupos C y D difieren en su desempeño; sin embargo, esto no se debe al estatus real del rasgo, sino que indica un sesgo en la prueba. Evidentemente, para determinar los sesgos, necesitamos información sobre el estatus real de los grupos en la variable, así como información sobre el desempeño en la prueba. Con este último ejemplo en mente, podemos relacionar la noción de neutralidad con el tratamiento formal de la validez, que presentamos en el capítulo 5. En particular, reintroducimos el concepto de varianza irrelevante para el constructo (véase figura 5-1). Una prueba específica busca medir un constructo; si una característica de la prueba interfiere con la medición exacta del constructo, esa característica introduce varianza irrelevante para el constructo. Regresemos al ejemplo de la persona con debilidad visual que trata de leer una prueba con un tamaño de letra de 10 puntos. La varianza asociada con el tamaño pequeño de la letra es irrelevante para lo que tratamos de medir y, por lo tanto, constituye un sesgo.

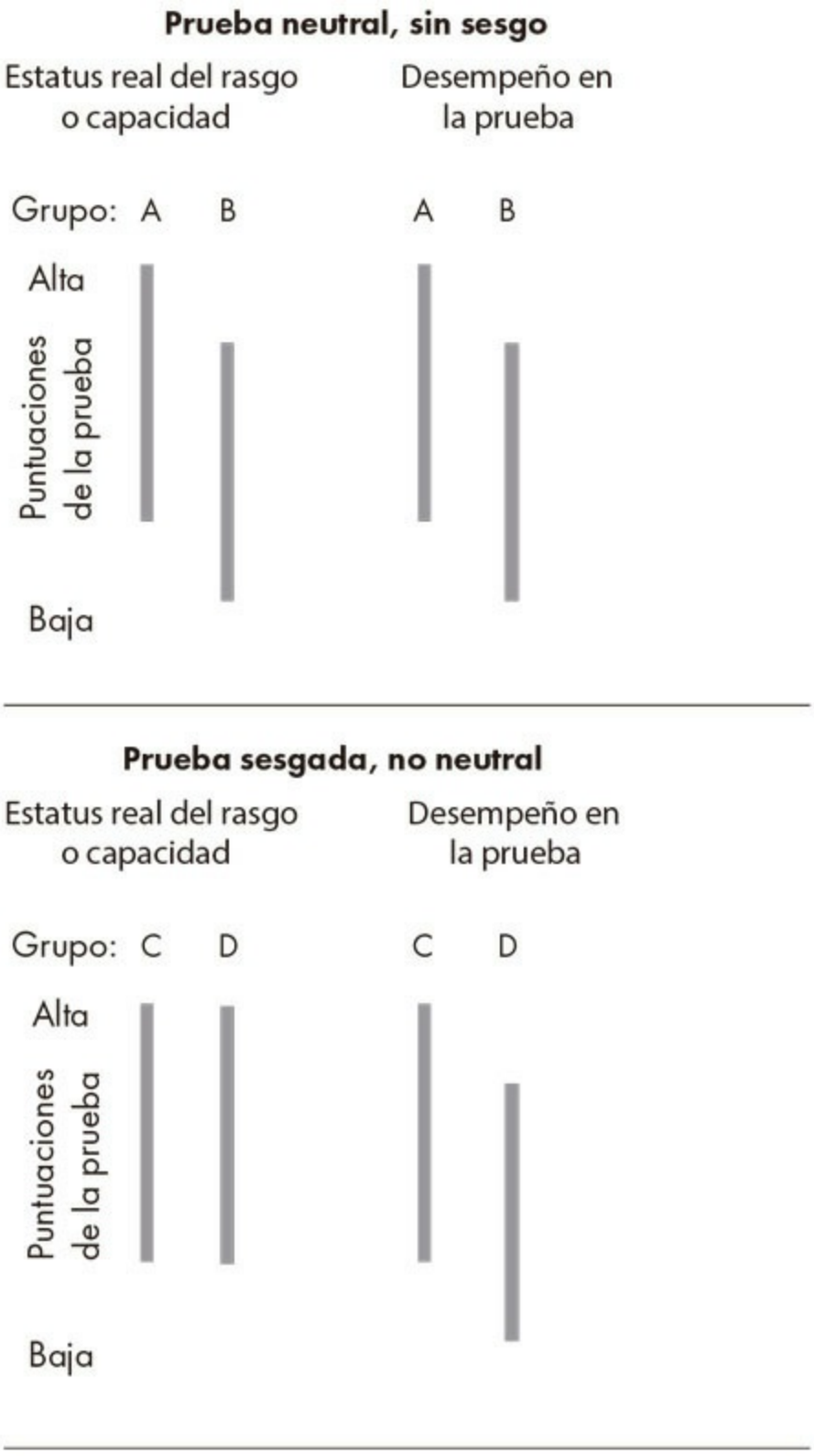


Figura 6-11. Ilustración del desempeño grupal en una prueba neutral y en una sesgada.

Una prueba que es neutral de acuerdo con el significado del *Standards* refleja el mismo constructo en todos

> los examinados, y sus puntuaciones tienen el mismo significado para todos los individuos de la población a la que está dirigida; una prueba neutral no da ventaja ni pone en desventaja a algunos individuos debido a características irrelevantes para el constructo que se quiere medir.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Los ejemplos que hemos citado hasta ahora provienen del campo de las pruebas de capacidad y aprovechamiento. De hecho, este campo ha sido el principal escenario para debatir la neutralidad de las pruebas. Sin embargo, las mismas nociones se aplican al campo de las pruebas de personalidad. Consideremos la medición de la introversión-extroversión, uno de los cinco grandes rasgos de personalidad. Una prueba que mida este rasgo podría elaborarse, estandarizarse y validarse en una cultura occidental de raza blanca y clase media. ¿La prueba funcionará igual para los examinados de una cultura nativa de América o para una del Medio Oriente? Aplicaremos a estas preguntas a los mismos tipos de análisis que empleamos con las medidas de capacidad o aprovechamiento. Por ejemplo, un estudio reciente investigó si el MMPI-2 funcionaba de manera adecuada (sin sesgos) en un grupo de una Antigua Orden Amish (Knabb, Vogt, & Newgren, 2011). En otro estudio se examinó si la Revised Children's Manifest Anxiety Scale [Escala de Ansiedad Manifiesta en Niños] funcionaba de manera adecuada (es decir, sin sesgo) con niños de Singapur (Ang, Lowe, & Yusof, 2011).

Métodos para estudiar la neutralidad de la prueba «163-165

¿Contamos con métodos para estudiar la neutralidad (sesgo) de las pruebas? Sí. Hay tres amplias categorías para estudiar la neutralidad y se usan mucho en la elaboración y el análisis de las pruebas. Algunos operan primordialmente durante la elaboración y otros, en cualquier momento en que se presenten preguntas acerca de su uso con diferentes grupos de examinados. Si tomamos cualquier número de una revista, como *Psychological Assessment*, es muy probable que encontremos varios estudios que analizan la neutralidad de una prueba para este o aquel grupo de examinados.

Revisión de panel

El primer método y el más sencillo para examinar la neutralidad de una prueba es la **revisión de panel**, la cual implica revisar los reactivos por su representatividad de varios grupos, que por lo común se refieren a grupos raciales, étnicos, culturales, socioeconómicos, con discapacidad y regionales. Los revisores tratan de reconocer reactivos que puedan contener frases o situaciones con diferente significado, connotación o grado de familiaridad para grupos específicos. Una prueba de lectura integrada en su totalidad por pasajes sobre futbol probablemente ponga en desventaja a las mujeres, mientras que un exceso de pasajes sobre danza ponga en desventaja a los hombres. Una prueba que emplee sólo escenarios de granja (equipo, animales, temporadas de crecimiento) probablemente pondría en desventaja a los habitantes urbanos, mientras que

un exceso de escenarios de los sistemas de transporte de la ciudad (metro, autobuses) ponga en esa situación a los habitantes de zonas rurales. Un reactivo que se refiera a comprar una botella de "pop" tiene perfecto sentido en la región del oeste medio, pero desconcertaría en la costa Este (donde pop es carbonato). La revisión de panel ayuda a reconocer palabras o situaciones que puedan tener un grado de familiaridad o significado diferentes para un grupo particular. Con esa información, el creador de la prueba trata de eliminar el material potencialmente problemático o equilibrar lo suficiente el material en los distintos grupos, de modo que, en promedio, ninguno quede en desventaja. El cuadro 6-12 contiene extractos de manuales de pruebas que se usan con mucha frecuencia; en ellos se habla sobre el proceso de revisión de panel. Estas revisiones ahora son una práctica casi universal en la elaboración de nuevas pruebas.

Cuadro 6-12. Afirmaciones muestra acerca de la revisión de panel de reactivos para evitar sesgos

"Panel de sesgos: Seis personas examinaron las tareas y los reactivos del MMSE-2 para detectar posibles sesgos o palabras ofensivas para grupos protegidos específicos. El panel incluyó un neuropsicólogo caucásico, un psicólogo hispano, un psicólogo asiático y tres profesionales no clínicos con diversos antecedentes étnicos: un caucásico, un afroamericano y un nativo americano." *Fuente*: Folstein, M. F. *et al*. (2010, p. 17). *Mini-Mental State Examination, 2nd edition*: *User's Manual*. Lutz, FL: PAR.

"Las formas de prueba de reactivos se enviaron al Bias Review Advisory Panel de educadores de minorías cuya principal preocupación fue eliminar cualquier posible fuente de sesgo... El panel reflejaba diversos antecedentes y representó a varias etnias, incluyendo afroamericanos, hispanos, asiático-americanos, americanos nativos y del Medio Este." *Fuente*: Otis, A. S. & Lennon, R. T. (2003, p. 13). *Otis-Lennon School Ability Test, Eight Edition: Technical manual*. San Antonio, TX: Harcourt Educational Measurement.

"Expertos en investigación transcultural y/o pruebas de inteligencia llevaron a cabo revisiones formales en tres ocasiones. Durante las fases iniciales del proyecto, todos… los reactivos fueron revisados por personas externas e internas para detectar posibles sesgos, obsolescencia cultural... Durante la fase de prueba, y otra vez durante la de estandarización, expertos en el contenido y en sesgos revisaron los reactivos e identificaron los que eran potencialmente problemáticos." *Fuente*: Wechsler, D. (2008a, p. 23). *Wechsler Adult Intelligence Scale – Fourth Edition: Technical and interpretive manual*. San Antonio, TX: Pearson.

El método de revisión de panel tiene dos inconvenientes. El primero se relaciona con el número de grupos representados: ¿cuántos incluir?, ¿quién podría faltar? En realidad, no hay límite para el número de grupos que podríamos identificar. Para tomar un ejemplo quizá tonto: ¿qué hay de la gente zurda de Wichita, Kansas?, ¿podrían estar en desventaja (o en ventaja) debido a un reactivo particular? Es evidente que este inconveniente exige al creador de la prueba usar su juicio.

El segundo inconveniente de la revisión de panel es que sus miembros se basan por completo en sus propias opiniones. Por un lado, un miembro puede identificar un reactivo o situación como problemáticos cuando, de hecho, puede no presentar ninguna ventaja o desventaja para ningún grupo. Por el otro, también puede pasar por alto un

reactivo o situación que sí lo son. De hecho, la investigación ha mostrado que los juicios de los miembros del panel acerca de qué reactivos podrían causar sesgos (desventajas) no son muy exactos (Engelhard, Davis, & Hansche, 1999; Engelhard, Hansche, & Rutledge, 1990; Plake, 1980; Sandoval & Miille, 1980). No obstante, la práctica continúa.

Funcionamiento diferencial de los reactivos (FDR)

La etapa de análisis de reactivos ofrece el contexto para estudiar el **funcionamiento diferencial de los reactivos**, al que, por lo general, nos referimos empleando su acrónimo **FDR**. El sesgo en los reactivos es un viejo término para este tema, pero en la literatura actual se prefiere el término, más neutral y quizá más descriptivo, *funcionamiento diferencial de los reactivos*. El FDR se refiere a la cuestión de si la prueba individual funciona de manera diferente para distintos grupos de examinados por razones distintas a las diferencias reales en el rasgo que se mide. De interés particular son las diferencias por raza, origen étnico y género; sin embargo, la cuestión básica puede referirse a cualquier comparación grupal, por ejemplo, entre personas de diferentes edades, estatura y lateralidad. Mientras que los procedimientos de revisión de panel eran únicamente cuestión de juicio, los del FDR buscan detectar sesgos mediante análisis estadísticos.

> ... se dice que el *funcionamiento diferencial de los reactivos* ocurre cuando examinados igualmente capaces difieren en sus probabilidades de responder de manera correcta un reactivo en función de la pertenencia a un grupo. El FDR se puede evaluar de varias maneras. Detectar el FDR no siempre indica sesgo en un reactivo, sino que es necesaria una explicación adecuada y sustancial del FDR para concluir que el reactivo tiene un sesgo.
>
> *Standards*... (AERA, APA, & NCME, 2013)

El punto más importante para comprender la discusión sobre el funcionamiento diferencial de los reactivos es que una simple diferencia en las dificultades del reactivo no necesariamente indica la presencia de un sesgo. Consideremos este caso. Examinamos el desempeño en un reactivo, el 23, de una prueba de aptitudes académicas; en el reactivo 23, 60% del grupo A[11] y 80% del grupo B respondieron de manera correcta, lo cual no significa que el reactivo tenga un sesgo en contra del grupo A. Supongamos que, en algún criterio externo del desempeño en el rasgo, determinamos que el grupo B, en efecto, tiene más del rasgo que el grupo A. Por ejemplo, podemos saber que el grupo A tiene un Grade Point Average (GPA) de 2.75, mientras que el grupo B tiene un GPA de 3.68; entonces, esperaríamos que el grupo B tuviera un mejor desempeño que el grupo A en el reactivo 23. Para nosotros, la diferencia entre 60% y 80% sería reflejo de una diferencia real en el rasgo; por el contrario, estaríamos desconcertados si los dos grupos tienen el mismo desempeño en el reactivo 23. Supongamos, por otro lado, que de acuerdo con el criterio externo, los grupos A y B son iguales en el rasgo que tratamos de medir: ambos

tienen un GPA promedio de 3.20; entonces, no esperaríamos que los grupos difieran en el reactivo 23 y, por lo tanto, no estaríamos inclinados a incluir este reactivo en la prueba final.

Se han propuesto numerosos métodos para estudiar el FDR. En este campo aún no se ha consolidado uno solo de estos enfoques. Una revisión de todos los métodos, o incluso de la mayoría de ellos, nos llevaría más allá de un texto introductorio como éste. Sin embargo, mencionaremos brevemente dos de los métodos del FDR más populares.

¡Inténtalo!

Si tienes acceso a un índice electrónico de literatura de investigación actual en psicología, educación o ciencias sociales, haz una búsqueda usando las palabras clave DIFFERENTIAL ITEM FUNCTIONING. Observa las diversas diferencias grupales que se estudian.

En el ejemplo anterior con los grupos A y B, establecimos la equivalencia de los grupos en el rasgo en términos de un criterio externo. En las aplicaciones más usuales del FDR, la equivalencia de los grupos se basa en la puntuación total de la prueba o en theta estimada (véase definición de theta en las pp. 49-50a»). Por lo común, el grupo más grande o mayoritario se denomina *grupo de referencia*, mientras que el más pequeño o minoritario se denomina *grupo focal*, es decir, el grupo en el que centramos nuestra atención. Entonces se examina el desempeño en los reactivos individuales. El **procedimiento Mantel-Haenszel** empieza dividiendo los grupos de referencia y focal en subgrupos con base en la puntuación total de la prueba. Pensemos en una prueba de 50 reactivos; dividamos la puntuación total en intervalos como se muestra en el cuadro 6-13. Entonces se determina el número de casos de los grupos de referencia y focal que tuvieron correcto o incorrecto cada uno de los reactivos. El estadístico Mantel-Haenszel deriva de este tipo de datos; dentro de un intervalo de puntuaciones, por ejemplo 31-40, los dos grupos se consideran iguales en el rasgo. La pregunta es si difieren en el desempeño en un reactivo individual. Los grupos completos, combinados a lo largo de todos los intervalos, bien pueden tener una diferencia media en el rasgo y aun así permitir un análisis de las diferencias en reactivos únicos. Por ejemplo, en el cuadro 6-13, el desempeño promedio es mayor en el grupo de referencia que en el grupo focal. Sin embargo, dentro de cualquier rango de puntuaciones, la razón entre respuestas correctas e incorrectas es casi la misma en los dos grupos.[12] El caso más obvio está en el rango de puntuaciones 11-20, donde la razón entre respuestas correctas e incorrectas es exactamente la misma en los dos grupos. Cincuenta por ciento de las puntuaciones del grupo focal está en este rango inferior, mientras que sólo 22% del grupo de referencia está en este rango. Sin embargo, el valor p del reactivo es exactamente .50 en los dos grupos y en ese rango.

Cuadro 6-13. Parte de los datos del análisis Mantel-Haenszel del FDR

Grupo de puntuaciones totales	1-10		11-20		21-30		31-40		41-50	
Desempeño en el reactivo 23[a]	+	−	+	−	+	−	+	−	+	−
Grupo de referencia	14	16	30	30	56	28	64	22	10	2
Grupo focal	10	12	20	20	15	8	10	4	5	1

[a] + = Correcto, − = Incorrecto.

Esta última descripción sugiere un análisis mucho más parecido a la curva característica del reactivo (CCR) descrita antes en este capítulo. De hecho, en la metodología de la TRR es una de las principales aproximaciones al FDR. En particular se desarrollan CCR de cada reactivo para los grupos que se están comparando. Los parámetros de estas curvas –dificultad, pendiente y adivinación– también pueden examinarse para determinar el FDR. La figura 6-12 ilustra esta aplicación. Las CCR que se superponen, o casi lo hacen, indican falta de FDR, como en el reactivo 19. El reactivo 27 muestra una CCR notablemente distinta, es decir, un FDR sustancial. El reactivo 36 ilustra un reactivo con un FDR notable en los niveles inferiores del rasgo (θ), pero sin diferencias en los niveles superiores. Podemos notar que estos análisis no dicen nada acerca del desempeño general de los dos grupos; puede ser que las puntuaciones promedio de los dos grupos sean diferentes por 20 puntos.

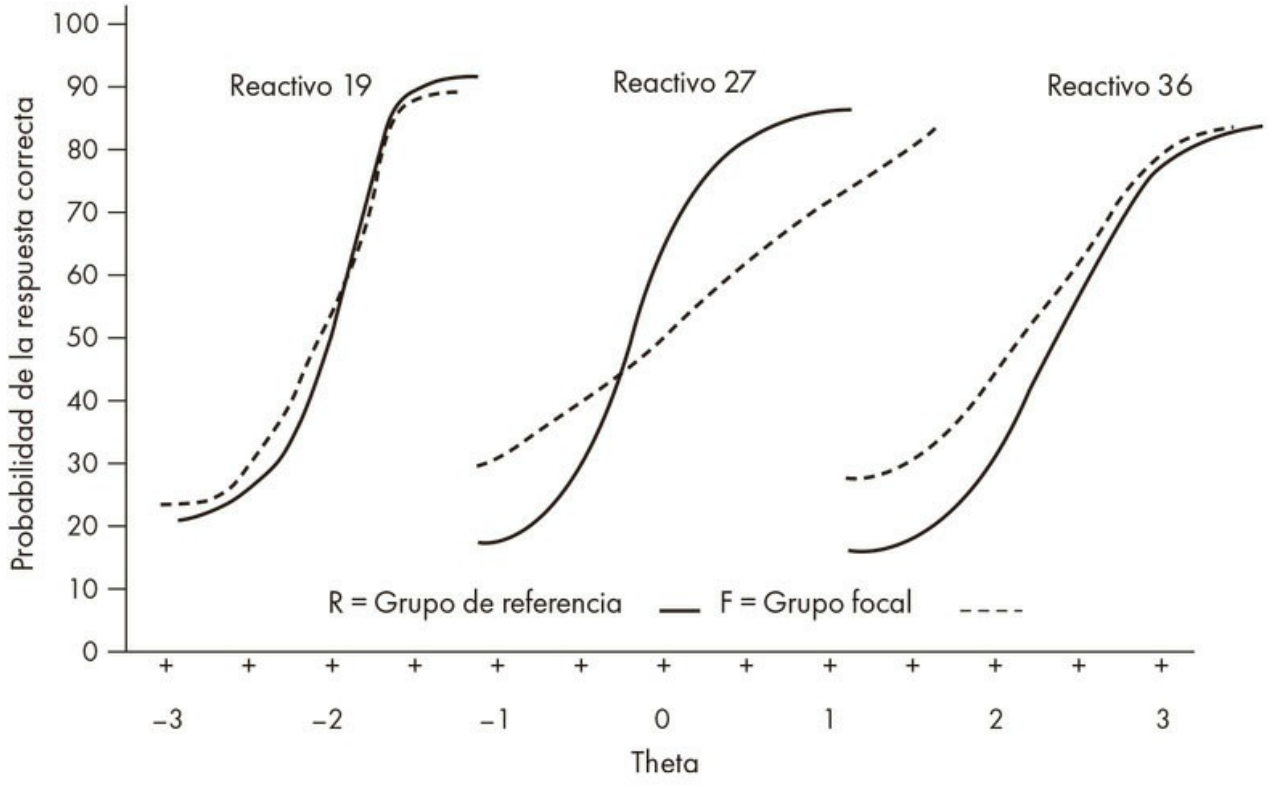


Figura 6-12. Análisis del FDR vía CCR de los dos grupos en tres reactivos.

Los análisis del FDR en el caso de diferencias raciales, étnicas y de género, por lo

general, se llevan a cabo en la etapa de análisis de reactivos durante la elaboración de la prueba. Sin embargo, muchos estudios del FDR se realizan después de que la prueba se publicó. Algunos de estos estudios aplican métodos del FDR nuevos o revisados, mientras que otros aplican las metodologías existentes a subgrupos nuevos. El número de subgrupos que se pueden analizar es prácticamente infinito.

Predicción diferencial

En el caso de las pruebas diseñadas para hacer predicciones, los métodos de validez de criterio, como se describieron en el capítulo 5, ofrecen un mecanismo importante para estudiar el sesgo de las pruebas. ¿Las pruebas funcionan de la misma manera con diferentes grupos, aun si los grupos varían en su desempeño promedio relacionado con diferencias reales en el rasgo subyacente? Una prueba sin sesgos debe producir predicciones igualmente buenas de varios grupos, lo cual no significa que el desempeño predicho en el criterio sea el mismo, sino que la predicción será igual de acertada para los dos (o más) grupos. En la siguiente discusión, siempre nos referiremos al contraste entre dos grupos, aunque la metodología con facilidad se extiende a comparaciones de cualquier cantidad de grupos.

> El término *sesgo predictivo* puede usarse cuando se encuentra evidencia de que existen diferencias en los patrones de asociaciones entre las puntuaciones de la prueba y otras variables en distintos grupos, lo que ocasiona preocupación por el sesgo en las inferencias basadas en las puntuaciones de la prueba. La predicción diferencial se examina usando el análisis de regresión. Un método examina la pendiente y las diferencias de la intersección entre dos grupos meta.
>
> *Standards*... (AERA, APA, & NCME, 2013)

En el contexto de la validez de criterio, en especial la validez predictiva, identificamos dos tipos de sesgo potencial: el sesgo de la intersección y el sesgo de la pendiente. Podemos notar que estos términos se relacionan con los dos parámetros de la ecuación de regresión (fórmula 5-1). El **sesgo de la intersección** significa que las intersecciones de las líneas de regresión difieren en los dos grupos. La figura 6-13 muestra un ejemplo de este sesgo; podemos notar que las pendientes de las líneas son iguales en los dos grupos. Como su nombre lo sugiere, el **sesgo de la pendiente** significa que las pendientes de las líneas de regresión difieren en los grupos. La figura 6-14 muestra un ejemplo de este sesgo.

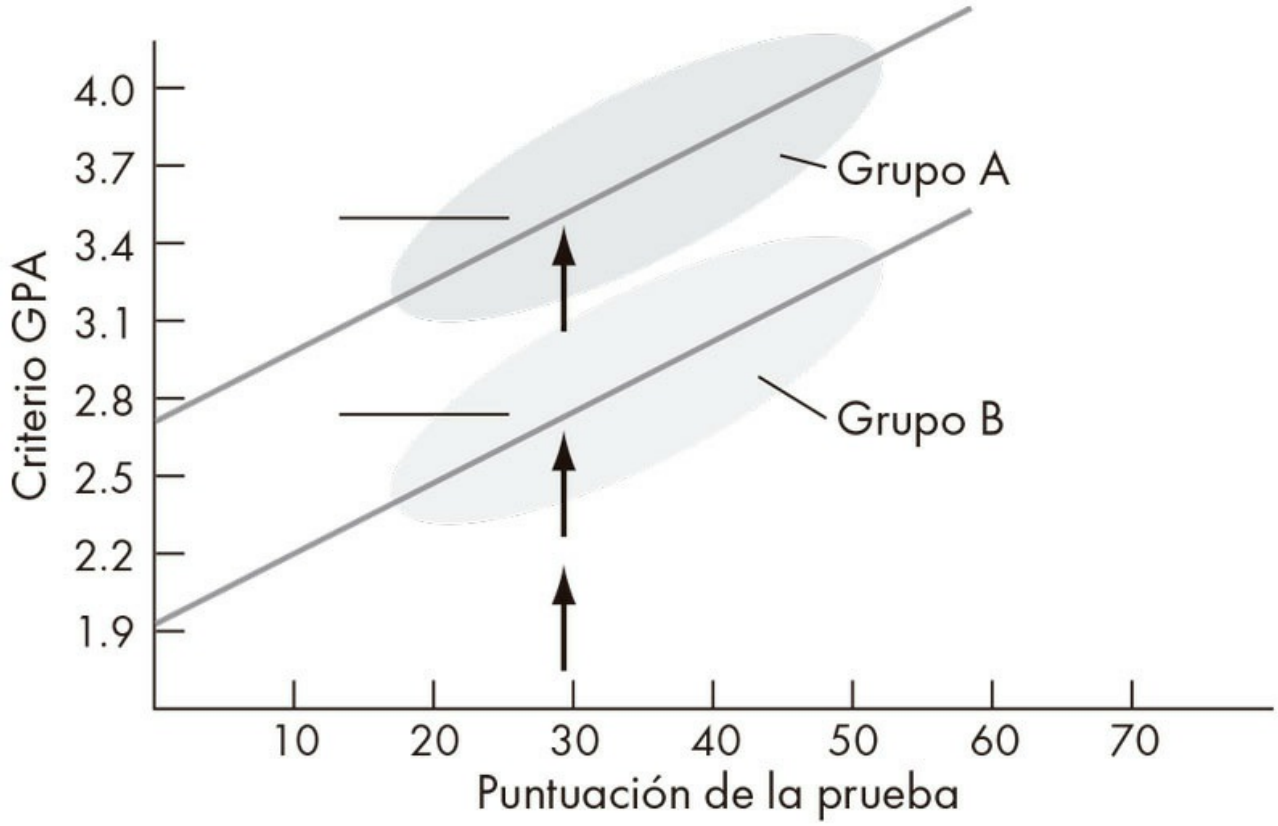


Figura 6-13. Ilustración del sesgo de la intersección.

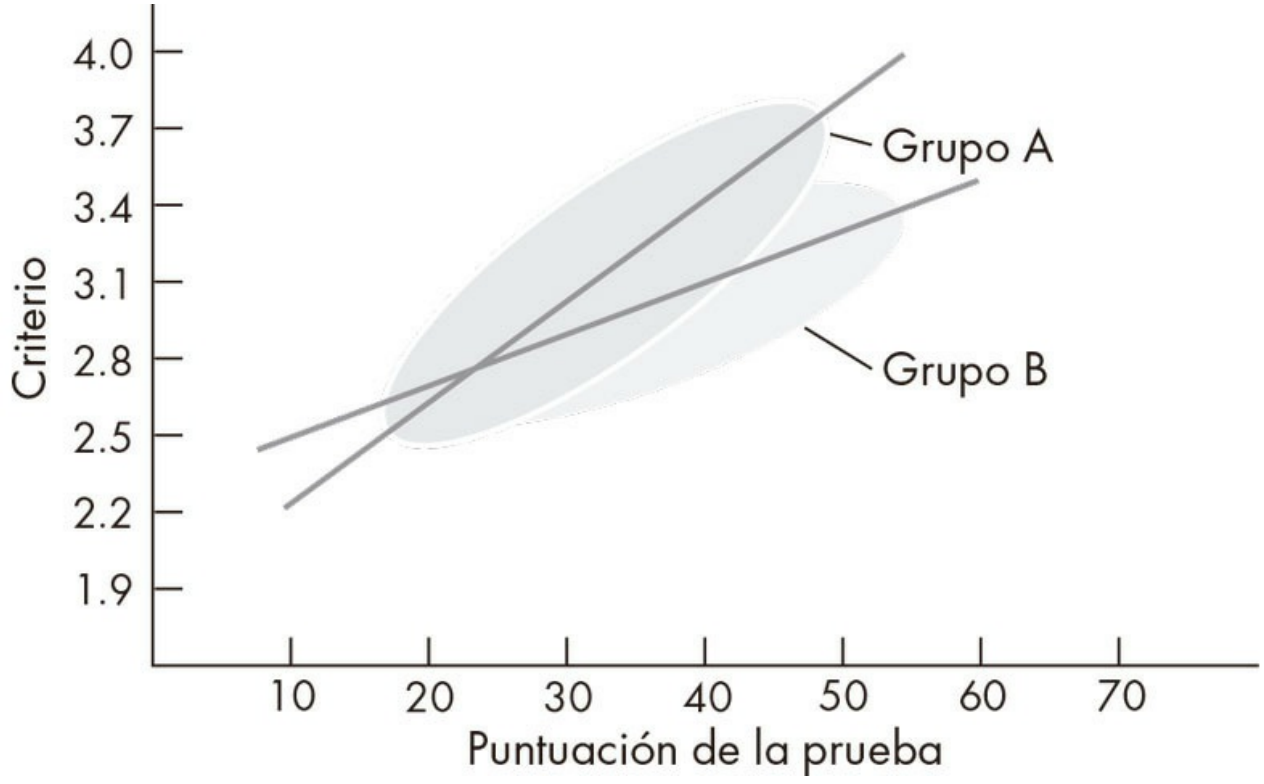


Figura 6-14. Ilustración del sesgo de la pendiente.

Consideremos estos dos conceptos con mayor detalle; usamos un ejemplo de predecir el GPA a partir de una prueba de admisión. En la figura 6-13, podemos notar cómo difieren las intersecciones; la del grupo A está alrededor del 2.8, mientras que la del grupo B, alrededor del 1.9. Por lo tanto, si una persona del grupo A obtiene una puntuación de 30 en la prueba de admisión, predeciríamos una puntuación del criterio (GPA) de cerca de 3.5; en cambio, si una persona del grupo B obtiene la misma puntuación en la prueba, 30, predeciríamos una puntuación del criterio de cerca de 2.8. Ésta es la situación más indeseable. Por lo común, la línea de regresión estaría determinada en los grupos combinados. Sin embargo, esta validez diferencial está escondida detrás de los resultados grupales globales. Podemos notar que la correlación entre la prueba y el criterio, en este caso, sería la misma en los grupos A y B. Éste es un ejemplo del sesgo de la intersección.

Ahora consideremos la figura 6-14, donde se presenta el caso del sesgo de la pendiente. Aquí la magnitud de la correlación es diferente en los dos grupos. La diferencia en las pendientes significa que habrá sobrepredicción en algunos casos del grupo A y en algunos del grupo B. De manera semejante, habrá subpredicción en algunos casos de cada grupo. También esta situación es muy indeseable. Éste es un ejemplo del sesgo de la pendiente; desde luego, es posible tener sesgo tanto de la pendiente como de la intersección.

La figura 6-15 muestra el caso en que hay una diferencia en el desempeño promedio de los dos grupos, pero no hay diferencias ni en la intersección ni en la pendiente. Podemos notar que una puntuación determinada, por ejemplo, 40, predice el mismo desempeño en el criterio sin importar la pertenencia grupal. Éste es el caso de los alumnos que estudiaron y los que no estudiaron para su examen de pruebas psicológicas. El grupo A estudió y el grupo B no estudió, por lo que el grupo A tiene puntuaciones superiores a las del grupo B. La prueba tiene la misma validez en la predicción del GPA.

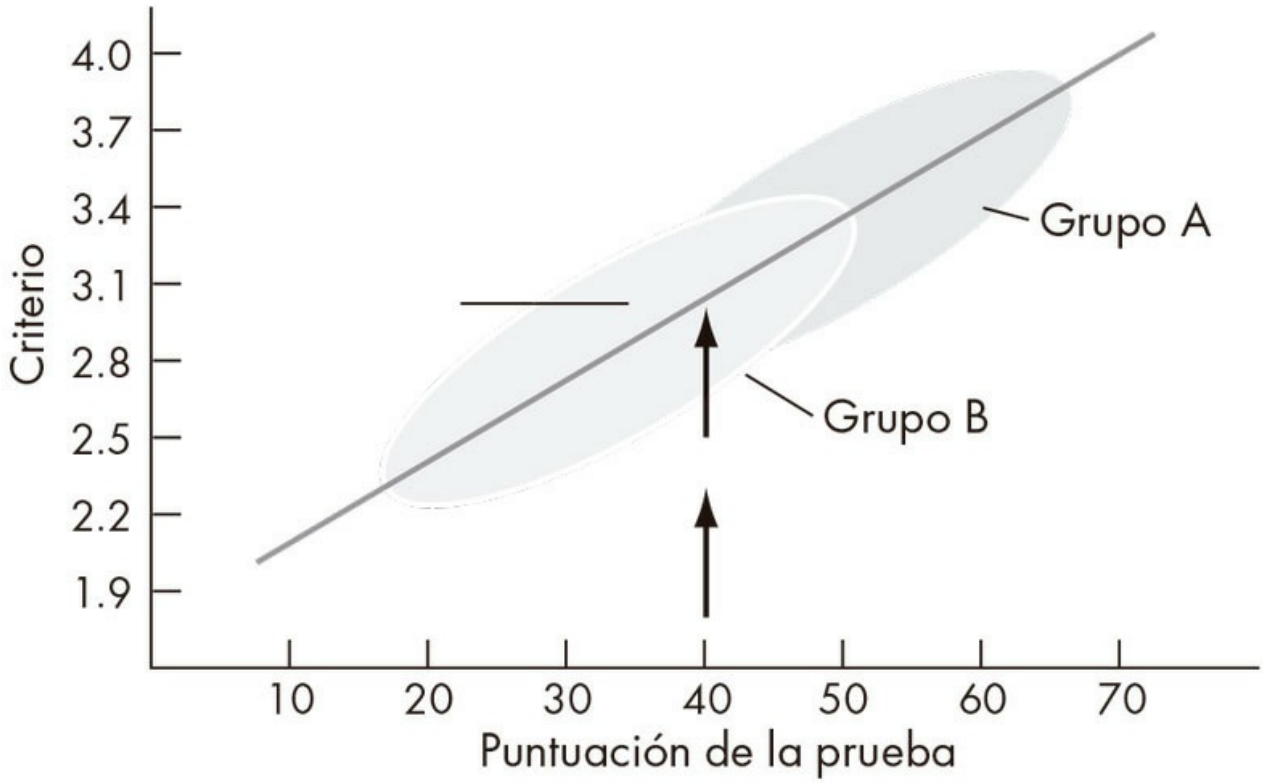


Figura 6-15. Ilustración de la ausencia de sesgo en la validez de criterio: pendientes e intersecciones iguales, pero diferencia en las medias.

Resumen de puntos clave 6-4

Tres métodos para estudiar la neutralidad de las pruebas

- Revisión de panel
- Funcionamiento diferencial de los reactivos
- Predicción diferencial

Invarianza de la medición

Hemos bosquejado tres métodos para estudiar la neutralidad de las pruebas: revisión de panel, funcionamiento diferencial de los reactivos y predicción diferencial. Estos son los tres métodos que se consignan en el *Standards*; todos buscan asegurar que una prueba mida un constructo específico (p. ej., comprensión de lectura, pensamiento creativo o depresión) de manera equivalente en diferentes grupos. Estos tres métodos son los que se encuentran comúnmente en la literatura profesional que se ocupa de la neutralidad. Sin embargo, otro grupo de técnicas, por lo general, clasificadas como análisis de la

invarianza de la medición, tiene exigencias más estrictas para la noción de medir un constructo de manera equivalente en diferentes grupos (p. ej., exige la demostración de la igualdad analítico-factorial, entre otras cosas). En la actualidad, las técnicas de invarianza de la medición no se usan mucho y van más allá de lo apropiado para este libro; sin embargo, merecen mencionarse porque pueden volverse más populares en el futuro. En Cheung y Rensvold (2002), Haynes, Smith y Hunsley (2011) y Vandenberg y Lance (2000) se pueden encontrar discusiones sobre los análisis de invarianza de la medición.

Adaptaciones y modificaciones

Una característica clave de una prueba psicológica o educativa es la *estandarización como procedimiento,* la cual implica usar los mismos reactivos, las mismas condiciones de aplicación, el mismo método de calificación... lo mismo de todo. ¿Pero qué pasa si esta "igualdad" significa que la prueba termina midiendo constructos diferentes en distintos grupos? La versión con letra grande de una prueba de comprensión de lectura usada con débiles visuales ofrece un ejemplo útil. Esta versión con letra grande no es exactamente la misma prueba (estandarizada) que la de letra normal; sin embargo, brindar una versión con letra grande ayuda a asegurar que la prueba mida el constructo meta (comprensión de lectura). Para los débiles visuales, la versión con letra normal es una prueba de agudeza visual, no de comprensión de lectura.

Nos referimos a los cambios en los procedimientos estandarizados de una prueba como **adaptaciones**. El término tiene su origen en los ajustes ambientales de personas con discapacidades físicas. Las rampas y los elevadores son adaptaciones evidentes para personas con silla de ruedas. Las adaptaciones en las condiciones de aplicación pueden implicar ediciones con letra grade de pruebas para personas con dificultades visuales. (Véase en el capítulo 16 la descripción de algunas leyes federales de EUA relacionadas con las adaptaciones.) Una adaptación particularmente controvertida es ampliar los límites de tiempo de las pruebas. La pregunta crucial es si una adaptación "iguala las condiciones" para una persona con una discapacidad o le da una ventaja no neutral sobre las personas que no reciben la adaptación. En términos técnicos, una adaptación en la aplicación debe volver igualmente aplicables la validez y las normas a los examinados con y sin discapacidad. Así, la versión de letra grande de la prueba de comprensión de lectura ayuda a "igualar las condiciones" entre personas con y sin debilidad visual. Una persona con visión normal no tendrá un mejor desempeño en la versión de letra grande que en la de letra normal. Así, la versión de letra grande parece una adaptación por completo razonable. Ahora consideremos el uso de un lector, es decir, una persona que lea la prueba a la persona con debilidad visual; esta adaptación bien puede cambiar la naturaleza de lo que se mide, pues ahora la prueba puede medir comprensión auditiva más que comprensión de lectura. Además, cualquier persona –con o sin debilidad visual– probablemente tendrá un mejor desempeño con un lector.

También consideremos la cuestión de ampliar los límites de tiempo en una prueba, adaptación que a menudo se propone en ciertos casos de problemas de aprendizaje. Si la

prueba es de poder puro (véase página 7a» del capítulo 1), dar tiempo adicional a alguien que lo pide debería ser aceptable; sin embargo, muy pocas pruebas son de poder puro. Los estudiantes con problemas de aprendizaje pueden tener un mejor desempeño si tuvieran más tiempo, pero lo mismo ocurriría con otros examinados (o tal vez no).

Numerosos estudios han examinado los efectos de varias adaptaciones sobre el desempeño. Claramente, aún tenemos mucho por aprender acerca de este tema. Varias fuentes han ofrecido útiles resúmenes de prácticas, regulaciones e investigación sobre adaptaciones de varios tipos de evaluaciones y grupos, por ejemplo, estudiantes con problemas de aprendizaje, estudiantes de inglés (Abedi, Hofstetter, & Lord, 2004; Camara, 2001; Nester, 1994; Pitoniak & Royer, 2001; Sireci, Li, & Scarpati, 2006; Thompson, Blount, & Thurlow, 2002; Thurlow, Elliott, & Ysseldyke, 1998; Thurlow & Ysseldyke, 2002; Willingham *et al*., 1988). Entre los puntos importantes que se abordan en estas fuentes se encuentra los siguientes. Primero, las prácticas y regulaciones están en un flujo continuo, lo que vuelve difícil hacer generalizaciones. Segundo, los efectos anticipados de una adaptación a veces se presentan, pero no siempre; por ejemplo, un tiempo ampliado (o cualquier otro cambio) puede no mejorar las puntuaciones del grupo meta; o el cambio puede mejorar las puntuaciones de todos los estudiantes, no sólo las del grupo meta. En esta última situación, el cambio no "iguala las condiciones", sino que, en realidad, ofrece una ventaja no neutral. Algunas adaptaciones funcionan como se piensa, pero no todas lo hacen. Tercero, está la cuestión de si las puntuaciones derivadas de una prueba aplicada con adaptaciones deben "señalarse", es decir, marcarse para indicar que la aplicación no fue estándar. Este es un tema de política en el que los profesionales de la evaluación no tienen una posición consensuada.

La práctica actual hace una distinción entre adaptación y modificación de una evaluación. En la adaptación, una persona responde, en esencia, la misma prueba que otras personas, pero con algunos cambios en las condiciones de aplicación, por ejemplo, una edición con letra grande o con límites de tiempo ampliados. Una modificación implica un intento de medir cierta habilidad o rasgo, pero con una metodología, en esencia, diferente. Por ejemplo, un maestro puede entrevistar a un estudiante con discapacidades múltiples para determinar elementos de la capacidad de solución de problemas matemáticos del alumno, porque ningún tipo de adaptaciones de la prueba regular escrita sería adecuado.

> Aunque los términos pueden tener significados diferentes bajo las leyes aplicables, como se usan en el *Standards*, *adaptación* denota cambios por los cuales la comparabilidad de las puntuaciones se mantiene, y *modificación* denota cambios que pueden afectar el constructo que mide una prueba.
>
> *Standards*... (AERA, APA, & NCME, 2013)

Como regla general, en la práctica profesional los resultados de una prueba con adaptaciones se consideran comparables con los resultados de la aplicación regular. De ahí que las normas puedan aplicarse a la versión adaptada, cuyos resultados pueden incluirse en los resúmenes de los grupos. En contraste, los resultados de una

modificación no se consideran comparables con la aplicación regular. Las normas regulares no deben aplicarse a la modificación, cuyos resultados no deben incluirse en los resúmenes de los grupos.

La literatura profesional hace una clara distinción entre adaptaciones y modificaciones. En realidad, las dos categorías probablemente representan un continuo subyacente de la desviación respecto de la aplicación estandarizada: desde la más trivial hasta la más extrema desviación que hace irreconocible la prueba original (estandarizada). Decidir cómo tratar las adaptaciones y/o modificaciones requiere un juicio cuidadoso y profesional. En "Guidelines for Assessment of and Intervention with Persons with Disabilities", de la *American Psychological Association* (2012), se puede encontrar una discusión de estos temas desde la perspectiva de la ética.

Algunas conclusiones tentativas

La investigación de la neutralidad de las pruebas, en particular de las de capacidad mental, se volvió muy activa a finales de la década de 1960. En ese tiempo, las pruebas de capacidad mental estuvieron sujetas a un examen, sobre todo por los sesgos raciales y étnicos, aunque el sesgo de género también fue un tema. Una serie de estudios llegaron a la conclusión de que las pruebas no evidenciaban una validez diferencial; los estudios también sugirieron que la estructura de las pruebas, como se determina, por ejemplo, mediante el análisis factorial, era sumamente similar en varios grupos. Quizá la referencia citada con mayor frecuencia es Jensen (1980), quien concluyó, después de una investigación exhaustiva sobre el tema, que "las pruebas estandarizadas más actuales de capacidad mental producían medidas sin sesgos en todos los segmentos angloparlantes nativos de la sociedad estadounidense actual, sin importar sexo, origen racial ni clase social. Las diferencias medias observadas en las puntuaciones de la prueba entre varios grupos, por lo general, no son un artefacto de las pruebas mismas, sino que son atribuibles a factores independientes, en términos causales, de las pruebas" (p. 740). Casi al mismo tiempo, Hunter, Schmidt y Hunter (1979) hicieron un resumen de 39 estudios sobre la posible validez diferencial por raza en el empleo de las pruebas. Concluyeron que "la verdadera validez diferencial probablemente no existe" (p. 721). Reynolds (1994) llegó a estas conclusiones: "Sólo desde mediados de la década de 19700 se han publicado investigaciones considerables en relación con el sesgo racial en las pruebas. En su mayor parte, esta investigación no ha apoyado la hipótesis del sesgo de la prueba, lo que revela que a) las pruebas psicológicas y educativas bien construidas y estandarizadas predicen el desempeño futuro de una manera, en esencia, equivalente en distintas razas de minorías étnicas nativas de EUA, b) la estructura psicométrica interna de las pruebas, en esencia, no está sesgada en favor de ninguna raza y c) el contenido de los reactivos en estas pruebas es casi igualmente apropiado para todos estos grupos" (p. 177). En una revisión más reciente, Reynolds y Ramsay (2003) concluyeron: "Existe el sesgo de las pruebas, pero es pequeño... A menudo sobreestima o sobrepredice el desempeño de las minorías, de modo que sus repercusiones sociales pueden ser muy diferentes de las que se le

suelen atribuir" (p. 87). Sin embargo, la cuestión de la validez diferencial necesita, de manera continua, tratarse conforme se elaboran nuevas pruebas o surgen preguntas acerca de la aplicabilidad de las pruebas existentes en diferentes subgrupos. La mayor parte de la investigación temprana sobre la neutralidad de las pruebas se concentró en las de capacidad y aprovechamiento, pero vemos un creciente número de estudios sobre las pruebas de personalidad.

Resumen

1. El primer paso en la elaboración de una prueba es redactar un propósito claro. En el enunciado se identifica la variable o constructo que se quiere medir y, por lo general, se incluye una referencia al grupo meta.
2. Después debe considerarse el diseño general de la prueba. Las consideraciones preliminares del diseño incluyen asuntos como la extensión de la prueba, el formato de reactivos, el número de puntuaciones, los procedimientos de calificación y la investigación de los antecedentes de la variable.
3. Entre las pruebas actuales, muchas surgieron para satisfacer algunas necesidades prácticas, otras, para propósitos teóricos. Gran parte del trabajo de elaboración de pruebas implica la adaptación o revisión de las pruebas actuales.
4. Los reactivos de respuesta cerrada, con diversos formatos, se usan mucho en las pruebas. El formato de opción múltiple es el más usado de este tipo de reactivos.
5. Los reactivos de respuesta abierta también se usan mucho en forma de ensayos, respuestas orales o evaluaciones de desempeño. Estos reactivos presentan desafíos especiales al calificarlos.
6. Hay varias sugerencias para redactar buenos reactivos, tanto de respuesta cerrada como de respuesta abierta.
7. El análisis de reactivos se refiere al conjunto de procedimientos para la prueba empírica y el tratamiento estadístico de los reactivos individuales. Hay tres fases: programa de prueba de reactivos, análisis estadístico y selección de reactivos.
8. Los estadísticos tradicionales de los reactivos incluyen el índice de dificultad del reactivo (p) y el índice de discriminación del reactivo (D o *r*).
9. En la metodología de la TRR, la curva característica del reactivo y sus parámetros, en especial los de dificultad y pendiente, son factores importantes al seleccionar reactivos.
10. El análisis factorial a veces se usa como técnica de análisis de reactivos.
11. Los datos del análisis de reactivos se usan, junto con otros criterios como las especificaciones del contenido, con el fin de elegir los reactivos para la prueba final.
12. Hay una relación entre el valor p del reactivo y su índice de discriminación máxima posible.
13. Las normas se desarrollan para la prueba final en el programa de estandarización. Diversos programas de investigación pueden tener lugar al mismo tiempo que el de estandarización.
14. La publicación final implica la prueba real, así como manuales, servicios de calificación y otros materiales complementarios.
15. La investigación sobre la prueba suele continuar después de su publicación. Parte de esta investigación será realizada por el autor y la editorial de la prueba, pero otros investigadores independientes también harán investigaciones sobre ella.
16. La neutralidad y el sesgo de las pruebas, términos alternos opuestos en su

connotación, tratan con la cuestión de si una prueba mide el mismo constructo subyacente en diferentes grupos.

17. Los métodos para estudiar la neutralidad de las pruebas incluyen la revisión de panel del contenido, el funcionamiento diferencial de los reactivos (FDR) y la predicción diferencial.

18. A veces, un examinado requiere adaptaciones en la prueba para ayudar a asegurar que ésta mida el mismo constructo en ese individuo que en los demás.

Palabras clave

adaptación
análisis de reactivos
calificación analítica
calificación automatizada
calificación holística
curva característica del reactivo (CCR)
cuestiones relacionadas con el diseño
diferencial semántico
dificultad del reactivo
discriminación del reactivo
escala de valoración gráfica
evaluación del desempeño
formato Likert
formulación del propósito
funcionamiento diferencial de los reactivos (FDR)
función informativa del reactivo
grupo superior e inferior
modelo de Rasch
modificación
neutralidad de la prueba
parámetro de adivinación
pendiente
procedimiento Mantel-Haenszel
programa de estandarización
prueba fácil
reactivos de respuesta abierta
reactivos de respuesta cerrada
revisión de panel
sesgo de la intersección
sesgo de la pendiente
sesgo de la prueba
sistema de puntos
tronco del reactivo
valor p

Ejercicios

1. Consulta los enunciados de los propósitos de las pruebas del cuadro 6-1. ¿Cómo podrías mejorar cualquiera de esos enunciados?
2. Planeas elaborar la mejor prueba del mundo, y la definitiva, de autoconcepto para estudiantes universitarios. Responde las siguientes preguntas acerca del diseño de tu prueba:

¿Cuántos reactivos tendrá? _____
¿Cuántas puntuaciones producirá? _____
¿Será de aplicación individual o grupal? _____
¿Cuánto tiempo tomará su aplicación? _____
¿Qué tipo de reactivos tendrá (opción múltiple, de respuesta abierta)? _____

3. Después de observar las directrices para redactar buenos reactivos, con el material de este capítulo:

- Escribe cinco reactivos de opción múltiple.
- Escribe cinco reactivos de verdadero-falso.
- Escribe cinco preguntas que se respondan con un ensayo.
- Pide a un compañero que critique tus reactivos.

4. Supón que quieres medir actitudes hacia la pena capital, es decir, el grado en que una persona está a favor o en contra de ella.

- Escribe cinco reactivos de tipo Likert sobre este tema.
- Crea cinco reactivos usando una escala de valoración gráfica.
- Pide a un compañero que critique tus reactivos.

5. Observa de nuevo los datos del cuadro 6-9.

- ¿Cuál es el valor p del reactivo 10?
- ¿Qué porcentaje de estudiantes del grupo inferior respondió el reactivo 23 de manera correcta?
- ¿Cuál es el reactivo más fácil del cuadro?
- ¿Cuál es la diferencia entre los grupos superior e inferior en el porcentaje de respuestas correctas en el reactivo 29?

6. Vuelve a ver el cuadro 6-13. Del reactivo 23, determina el valor p del grupo completo de referencia y luego el del grupo focal.
7. Revisa los datos del análisis de reactivos que aparecen abajo. D es el índice de discriminación y p es el índice de dificultad.

¿Cuáles son los dos reactivos que eliminarías si quisieras hacer una prueba final más

fácil?
¿Cuáles son los dos reactivos que eliminarías si quisieras hacer una prueba final más difícil?
¿Cuáles son los dos reactivos que eliminarías para aumentar la consistencia interna de la prueba?

Reactivo	p	D
1	.60	.20
2	.75	.25
3	.55	.05
4	.90	.15
5	.35	.30
6	.65	.35
7	.60	.40
8	.40	.15
9	.80	.25
10	.85	.10
11	.70	.30
12	.50	.25

8. Observa de nuevo los estadísticos de reactivos del ejercicio 7. Supón que estás creando una prueba de cinco reactivos y seleccionaste los reactivos 2, 4, 9, 10 y 11. Supón que los estadísticos de los reactivos se basan en una muestra representativa. ¿La distribución de las puntuaciones de la prueba se parece más a la de la prueba A o a la de la prueba B de la figura 6-10?
9. Accede a una reseña, en formato electrónico o impreso, de cualquier prueba en una edición reciente del MMY de Buros. ¿Qué dice la reseña acerca del programa de estandarización de la prueba? ¿De qué tamaño fue el grupo de estandarización? ¿Era representativo de la población a la que está dirigida la prueba?
10. Con los datos de reactivos del apéndice D, genera los estadísticos p y D con ayuda de algún programa de cómputo.
11. Supón que estás construyendo una prueba de comprensión de lectura como auxiliar en la evaluación de problemas de lectura. Quieres tener una *revisión de panel* del contenido de tus reactivos. Haz una lista de los tipos de personas que quieres que formen parte del panel.
12. Calcula el valor p global del reactivo del cuadro 6-13 en los grupos de referencia y focal. (Para obtener p en uno de los grupos, suma todas las respuestas "+" y divide el resultado entre la suma de todas las respuestas "+" y "–" del grupo.) ¿Cómo se comparan los valores p globales? ¿Qué dice esto acerca de los grupos de referencia y focal? Ahora obtén el valor p dentro de cada rango de puntuaciones de los grupos de referencia y focal por separado. Divide el número de respuestas "+" entre la suma de

las respuestas "+" y "–". ¿Qué te dice esta información?

13. Accede a la hoja de cálculo "Generador de CCR.xlsx", disponible en el apéndice D, junto con el documento de "Instrucciones para usar el Generador de CCR". La hoja de Excel te permite "jugar con" los valores en el modelo de la TRR de tres parámetros. Tú puedes construir tu propia CCR, como se describió en este capítulo.

14. Entra al sitio http://www.metheval.uni-jena.de/irt/VisualIRT.pdf, que contiene una gran cantidad de *applets* que te permiten variar los parámetros para diferentes funciones de la TRR. Será muy divertido.

Notas

1 Esta lista difiere ligeramente de la que aparece en el *Standards for Educational and Psychological Testing*, donde se incluyen cuatro pasos. Los primeros dos son, en esencia, los mismos que presentamos aquí. Nuestros pasos 3 y 4 aparecen combinados en el *Standards* en uno solo; nosotros los separamos, porque son muy diferentes en términos lógicos y cronológicos. De manera inexplicable, el *Standards* no incluye nuestro paso 5 a pesar de que hace una remisión del capítulo de desarrollo de pruebas a los capítulos sobre normas, validez y confiabilidad. El paso 4 del *Standards* es el mismo que nuestro paso 6.

2 Como señalamos en el texto, hay numerosos términos alternativos para denominar los formatos de respuesta abierta y de respuesta cerrada.

3 Técnicamente, el formato Likert se refiere al método de construir la escala entera; sin embargo, es común referirse al formato de respuesta mismo como formato Likert. En el capítulo 15 se puede consultar una discusión más amplia sobre este tema.

4 Algunas veces se hace referencia *al* diferencial semántico como si fuera una prueba específica. Sin embargo, como se señala en la obra clásica sobre este tema (Osgood, Suci, & Tannenbaum, 1957), el diferencial semántico se refiere a una técnica general, no a una prueba específica.

5 En diferentes aplicaciones, la puntuación total de la prueba puede definirse de varias maneras distintas. Por ejemplo, puede basarse en todos los reactivos de la prueba excluyendo el que se está analizando, o en todos los reactivos de una subprueba perteneciente a una batería más grande.

6 Aunque 27% parece una cifra extraña, hay una buena razón para usarla. Cuando se contrastan grupos, queremos optimizar dos condiciones que, por lo general, trabajan una contra otra. Por un lado, queremos que los grupos sean tan diferentes como sea posible; de acuerdo con este principio, un contraste entre 5% o 10% superior e inferior sería mejor que, digamos, entre 50% superior e inferior. Por otro lado, para obtener datos estables, queremos que los grupos sean tan grandes como sea posible; de acuerdo con este principio, grupos con 50% superior e inferior serían preferibles. En un famoso análisis publicado en 1928, pero después corregido, Truman Kelley (1939) mostró que la solución óptima a este problema era usar 27% superior e inferior; de ahí que esta cifra se haya convertido en el "estándar industrial" para dividir los grupos. A menudo se usan 25% o 33% superior e inferior como aproximaciones razonables a 27%, con los beneficios adicionales de permitir un análisis de uno o dos grupos intermedios del mismo tamaño que los grupos de los extremos.

7 Puede haber una asíntota inferior por razones distintas a la de la adivinación al azar. De ahí que la asíntota inferior a veces se denomine parámetro de seudoadivinación.

8 En algunas aplicaciones, el análisis de reactivos y la estandarización irán acompañados de un programa único de investigación. Sin embargo, ésta no es la práctica habitual. Combinar con éxito el análisis de reactivos y la estandarización requiere un cuidado excepcional y maestría (y, quizá, suerte).

9 Hay un límite a esta generalización. Si el rasgo que tratamos de medir tiene una definición muy restringida y

está muy marcadamente focalizado, entonces son deseables los índices de discriminación muy altos. Si el rasgo tiene una definición más amplia, difusa y compleja, entonces son deseables los índices de discriminación moderados (pero aún claramente positivos). En la práctica, casi nunca hay que preocuparse por tener índices de discriminación que sean demasiado altos.

10 Análisis como éstos suponen que los examinados adivinan al azar siempre que sea posible. De hecho, a menudo no adivinan al azar, y pueden no adivinar en absoluto cuando no saben la respuesta a la pregunta.

11 Una vez más, la terminología básica viene del campo de las pruebas de capacidad y aprovechamiento, pero los conceptos se aplican igualmente a las medidas de personalidad, intereses y actitudes. De ahí que aquí decimos "responde de manera correcta", pero también podríamos decir sólo "responde de manera afirmativa" o "responde Sí". Podemos notar que la afirmación del *Standards* se refiere a una "capacidad igual", pero, por extensión, significa "igual en el rasgo".

12 Técnicamente, el procedimiento Mantel-Haenszel realiza un análisis de chi-cuadrada con los datos. Los intervalos de puntuaciones, por lo común, serían más restringidos que los del cuadro 6-13 y pueden estar en intervalos de unidades a lo largo de toda la distribución.

SEGUNDA PARTE

La primera parte de este libro se concentró en los principios y procedimientos fundamentales aplicables a todo tipo de pruebas. En la segunda, seguimos con el examen de tipos específicos de pruebas, tal como las dividimos en varias categorías principales en el capítulo 1. Después de la introducción a las teorías de la inteligencia, que presentamos en el capítulo 7, en los capítulos 8 al 11 abordamos las pruebas que pertenecen primordialmente al dominio cognitivo, mientras que en los capítulos 12 al 15 tratamos las que pertenecen al campo no cognitivo. En cada capítulo de esta segunda parte se bosquejan los principales usos de las pruebas de cada categoría; luego, se identifican las pruebas de uso más frecuente en esa categoría y, después, se tratan de manera detallada algunas de las pruebas más usadas. Dentro de la mayoría de las categorías principales de pruebas, hay literalmente cientos de ejemplos que podrían introducirse; de acuerdo con nuestra orientación práctica, hemos decidido, en general, presentar sólo las pruebas más usadas. En definitiva, evitamos la detestable práctica de catalogar numerosas pruebas haciendo la descripción más escueta posible de ellas. Además, una lista así la podemos obtener de las fuentes de información que mencionamos en el capítulo 2. Por último, cada capítulo o sección principal de un capítulo concluye con algunas observaciones o preguntas acerca del estado actual de las pruebas dentro de la categoría que se está tratando.

CAPÍTULO 7
Inteligencia: teorías y temas

Objetivos

1. Describir el significado del término inteligencia y sus correlatos prácticos.

2. Describir características clave de las siguientes teorías de la inteligencia:

"g" de Spearman

Capacidades mentales primarias de Thurstone

Modelo jerárquico

Teorías del desarrollo

Modelos del procesamiento de información y biológicos

3. Describir la relación entre teorías y pruebas de inteligencia.

4. En cada una de estas comparaciones grupales, identificar los principales resultados

de la investigación: sexo, edad, estatus socioeconómico y grupos raciales/ étnicos.

5. Resumir los principales resultados de los estudios sobre el factor hereditario en la inteligencia.

Inteligencia: áreas de estudio

El estudio psicológico de la inteligencia abarca cuatro áreas amplias de interés interrelacionadas, pero distintas. La primera es la de las teorías acerca de la naturaleza de la inteligencia. La segunda es la metodología, tanto teórica como aplicada, de la medición de la inteligencia. La tercera es el área de las diferencias grupales en relación con la inteligencia: por edad, género, nivel socioeconómico, grupos raciales/étnicos. La cuarta es la cuestión de las influencias hereditarias y ambientales en el desarrollo de la inteligencia. Nuestra principal preocupación en este libro, obviamente, es la del segundo punto: la medición de la inteligencia. Los capítulos 8 y 9 están dedicados por completo a este tema; pero antes necesitamos estudiar el primer punto, el de las teorías de la inteligencia, debido a su interacción con los procedimientos de medición. Las buenas teorías ayudan a dirigir la medición y ésta ayuda a estimular y asegurar los desarrollos teóricos. En este libro, en realidad no necesitamos estudiar los puntos tercero y cuarto; no obstante, el estudiante de pruebas psicológicas suele tener un fuerte interés por los asuntos de las diferencias grupales y las influencias hereditarias y ambientales sobre la inteligencia. Por ello, en la parte final de este capítulo, haremos un repaso general de estos temas, pero sin intentar hacer un tratamiento exhaustivo de ellos. Bosquejaremos las principales conclusiones de estos temas y citaremos las fuentes donde se pueden consultar. Antes de retomar cada uno de ellos, debemos hacer una pausa para considerar el significado del término inteligencia, los otros nombres con que se le menciona y los hallazgos generales acerca de las correlaciones de la inteligencia.

Significado de inteligencia

¿Qué es la inteligencia? Los psicólogos se han ocupado de esta pregunta desde los inicios del siglo XX, cuando Spearman teorizó por primera vez sobre su estructura y Binet introdujo la primera medición práctica de ella. No es raro encontrar afirmaciones de que no sabemos qué significa el término inteligencia. En realidad, entre quienes estudian este tema con diligencia, hay un gran acuerdo acerca del significado de este término, al menos en la especie humana y en las culturas occidentales. De hecho, quizá es mayor el acuerdo acerca del significado de inteligencia que de otros constructos psicológicos como "personalidad". A decir verdad, no hay un acuerdo universal sobre la definición de inteligencia, y podemos encontrar autores que dan definiciones inusuales de este término, pero entre los psicólogos que estudian con regularidad esta área hay un sorprendente consenso en relación con el significado del término que ha durado años.

Resumen de puntos clave 7-1

Las cuatro áreas principales del estudio psicológico de la inteligencia
1. Teorías de la inteligencia

2. Medición de la inteligencia
3. Diferencias grupales en relación con la inteligencia
4. Influencias hereditarias y ambientales sobre la inteligencia

Varios estudios ofrecen listas útiles de los elementos (o manifestaciones) de la inteligencia. Una serie de artículos del Journal of Educational Psychology (1921), en los que aparecen autores legendarios como Terman, Thorndike y Thurstone, abordó, en parte, la definición de inteligencia.[1] Sesenta y cinco años después, Sternberg y Detterman (1986) realizaron un esfuerzo similar con otro panel de distinguidos expertos. Aunque sus esfuerzos por resumir los puntos de vista expresados en los dos simposios fueron difusos, en el mejor de los casos, el principal punto de acuerdo en ambos fue la concentración en el "razonamiento abstracto, representación, solución de problemas y toma de decisiones" (Sternberg & Detterman, p. 158). Snyderman y Rothman (1987) encuestaron a 661 psicólogos y especialistas en educación elegidos de manera deliberada para representar a los expertos del área de la inteligencia. Más de 95% de los encuestados estuvo de acuerdo con esta descripción de inteligencia: pensamiento o razonamiento abstracto, capacidad para solucionar problemas y capacidad para adquirir conocimientos nuevos; y entre 70 y 80% estuvo de acuerdo en que la inteligencia incluía la memoria, la adaptación al ambiente y la velocidad mental. Neisser et al. (1996, p. 77) resumieron las ideas de una comisión nombrada por la American Psychological Association para desarrollar una declaración de consenso sobre una gran cantidad de temas relacionados con la inteligencia, a la cual se hace referencia como "la capacidad de comprender ideas complejas, adaptarse de manera eficaz al ambiente, aprender de las experiencias, aplicar varias formas de razonamiento y superar obstáculos por medio del pensamiento". Un panel de 52 investigadores de vanguardia en el área de la inteligencia aprobaron esta definición: "Inteligencia es una capacidad mental muy general que, entre otras cosas, implica la capacidad para razonar, planear, resolver problemas, pensar de manera abstracta, comprender ideas complejas, aprender con rapidez y aprender de la experiencia" (Gottfredson, 1997, p. 13). En un intento por actualizar el resumen de Neisser et al., Nisbett et al. (2012) adoptaron explícitamente la definición de Gottfredson; estas fuentes con frecuencia hacen hincapié en que la inteligencia no es la sola maestría en un campo. La figura 7-1 resume los términos que se usan, por lo común, en la definición de inteligencia.

Figura 7-1. Términos que se usan comúnmente en la definición de "inteligencia".

Correlatos de la inteligencia con el mundo real

¿La inteligencia, como se acaba de definir, hace alguna diferencia en la vida de las personas? Responder a esta pregunta requiere de algunos análisis empíricos, y de cierta perspectiva de los resultados empíricos y cierta perspectiva de la vida misma. Tratar con este primer tema es bastante fácil; tratar con el segundo es un poco más difícil, pero aún está dentro del reino donde los conceptos psicométricos y estadísticos son aplicables; tratar con el tercer tema nos lleva a regiones de la filosofía y la religión donde nuestras competencias merman.

Un gran número de estudios ha abordado la primera pregunta, por lo común, mostrando la correlación entre el desempeño en un prueba de inteligencia y apoyándose en algún criterio externo o mostrando diferencias en los promedios grupales. Los criterios externos de interés incluyen el aprovechamiento académico –definido por índices como calificaciones, pruebas de aprovechamiento o desarrollo en niveles superiores de educación–, desempeño en un trabajo –definido por índices como valoración de los supervisores y promociones–, estatus socioeconómico –definido por índices como ingreso y prestigio ocupacional– y cuestiones generales de "calidad de vida" –definida por índices como estado de salud, transgresiones a la ley e integridad familiar. En general, la inteligencia se relaciona positivamente con estos criterios externos; las correlaciones con el aprovechamiento académico son las más altas y van de .80, cuando el criterio es una prueba de aprovechamiento, hasta .50, cuando el criterio son las calificaciones. Las correlaciones con el desempeño en el trabajo, en el caso de una amplia gama de ocupaciones, pero en especial cuando se trata de niveles profesionales o de habilidades, tienden a variar entre .30 y .60 en diferentes estudios. Las correlaciones (o diferencias grupales) con varios aspectos relacionados con la salud y la calidad de vida varían de manera considerable, pero en una clara dirección positiva. En Deary y Batty (2011), Gottfredson (2004), Kuncel, Hezlett y Ones (2004) y Schmidt y Hunter (2004), así como en libros enteros dedicados al tratamiento general de la inteligencia, como el de Hunt (2011, en especial el capítulo 10) y Mackintosh (1998), podemos encontrar

resúmenes de la investigación pertinente sobre estas cuestiones.

Respecto de la investigación correlacional, el lector debe estar enterado de que las técnicas que se emplean usualmente incluyen varias correcciones de atenuación (confiabilidad imperfecta) que presentamos en el capítulo 5 y correcciones de la restricción de rango que presentamos en el capítulo 4. Estas correcciones no son sólo tecnicismos etéreos, sino que se usan en la realidad.

Hay un gran consenso en relación con la magnitud general de la relación entre la inteligencia y los tipos de criterios que acabamos de bosquejar. Las opiniones empiezan a dividirse al interpretar los datos. Para un intérprete, una correlación de .50 puede reflejar una relación "fuerte", incluso excepcionalmente fuerte si se trata de los tipos de relación que se suelen encontrar en el mundo de la conducta. Para otro intérprete, esa misma r = .50 no es tan impresionante; después de todo, sólo explica 25% de la varianza. Del mismo modo, una correlación de, digamos, .30 entre una prueba de inteligencia y algún índice de desempeño en el trabajo puede justificar la afirmación de que la inteligencia es "el predictor más poderoso que se conoce" del desempeño en el trabajo, pero al mismo tiempo r = .30 simplemente no es un predictor muy poderoso. Estos ejemplos sencillos nos recuerdan que los hechos desnudos acerca de las correlaciones de la inteligencia están sujetos a interpretaciones variables. Por un lado, la inteligencia, tal como se mide mediante las pruebas que se describen en los siguientes dos capítulos, tiene claras relaciones dignas de atención con una gran variedad de variables que se suelen considerar importantes en la vida. Cualquiera que niegue eso no está poniendo atención a los hechos. Por otro lado, la inteligencia no nos cuenta la historia completa acerca de nada; cualquiera que piense esto tampoco está poniendo atención a los hechos. Otros factores que hacen la diferencia incluyen la educación, experiencia, personalidad, motivación, esfuerzo y puro azar.

¿Cómo llamarlas?

Quizá más que la de cualquier otro tipo, las pruebas que consideramos en los siguientes capítulos parecen estar en una crisis perpetua de identidad. Los psicólogos se sienten satisfechos llamando prueba de personalidad a una prueba de personalidad y prueba de aprovechamiento a una prueba de aprovechamiento. Sin embargo, sienten ansiedad y ambivalencia al tener que nombrar las pruebas que se abordan en los capítulos 8 y 9.

En los primeros días, los nombres estándar de estas pruebas eran de inteligencia, de capacidad mental o de aptitud; de hecho, algunas pruebas mantienen esos nombres. Sin embargo, una gran cantidad de términos alternativos ha surgido en los años recientes, y esto se debe a dos razones. Primero, las teorías acerca de la estructura de la inteligencia humana han evolucionado con el tiempo; la siguiente sección examina este tema. De ahí que los autores de las pruebas hayan buscado términos que se apeguen con más claridad al propósito específico de la prueba. Por ejemplo, el *Otis-Lennon Mental Ability Test* [Prueba de capacidades mentales Otis-Lennon] (sucesor del *Otis Intelligence Scale*) ahora se llama *Otis-Lennon School Ability Test* [Prueba de capacidades escolares Otis-

Lennon]. Es decir, más que referirse a la "capacidad mental" exacta, la prueba hace hincapié en que intenta medir las capacidades que se necesitan en la escuela. Segundo, ha habido un gran temor a que la gente piense que la inteligencia o la aptitud es innata o exclusivamente hereditaria. Para contrarrestar esta idea, los autores de pruebas han elegido términos que subrayan que la prueba mide capacidades desarrolladas, no aptitudes innatas. Por ejemplo, el *Scholastic Aptitude Test* [Prueba de aptitudes escolares] se convirtió en el *Scholastic Assessment Test* [Prueba de evaluación escolar], y después simplemente en SAT. Así, ahora encontramos desconcertantes combinaciones de términos como "evaluación cognitiva", "capacidades diferenciales" o sencillamente "vocabulario" en los títulos de las pruebas generalmente consideradas como medidas de inteligencia o capacidad mental. En éste y el siguiente capítulo, usamos los términos inteligencia y capacidad mental asumiendo que el lector tendrá en mente estas últimas advertencias.

Teorías de la inteligencia

Las teorías acerca de la naturaleza de lo que tratamos de medir tienen un papel más destacado en el área de las pruebas de inteligencia que en cualquier otra área. De ahí que, antes de describir pruebas particulares, resumimos las principales teorías de la inteligencia –sólo lo suficiente para brindar los antecedentes necesarios para comprender el desarrollo y aplicación de las pruebas de inteligencia. En Hunt (2011) y en varios artículos del *Cambridge Handbook of Intelligence* [Manual Cambridge de inteligencia] (Sternberg & Kaufman, 2011) se puede encontrar una discusión minuciosa de estas teorías.

En los manuales de las pruebas de inteligencia abundan las referencias a las teorías de la inteligencia. En el cuadro 7-1 se pueden leer afirmaciones muestra extraídas de dichos manuales. Necesitamos estar familiarizados con las teorías para comprender las pruebas; por ejemplo ¿qué es la "teoría actual de la inteligencia" que se cita en el manual de WAIS-IV? ¿Qué es un factor de orden superior? ¿Qué es la "estructura jerárquica" que se menciona en el manual del Otis-Lennon? ¿Qué significa la "g" que se menciona en el Otis-Lennon y en la Stanford-Binet? Estas citas hacen referencia a elementos de distintas teorías de la inteligencia.

Cuadro 7-1. Afirmaciones muestra que hacen referencia a las teorías de la inteligencia en los manuales de las pruebas

Del manual de la Escala Wechsler de Inteligencia para Adultos – IV: "Como WISC-IV, WAIS-IV ofrece una medida del funcionamiento intelectual general (CI Total) y cuatro puntuaciones índice... El nuevo marco se basa en la teoría actual de la inteligencia y se apoya en los resultados de la investigación clínica y el análisis factorial..." (Wechsler, 2008a, p. 8)
Del manual técnico del Otis-Lennon School Ability Test, octava edición: "El marco teórico más satisfactorio de la serie del OLSAT es la estructura jerárquica de las capacidades humanas... [el] 'factor general' (g) se ubica en la parte más alta de la jerarquía. Un nivel abajo están los dos principales factores grupales... Inmediatamente abajo de éstos se encuentran los factores grupales menores y los factores específicos." (Otis & Lennon, 2003, p. 5)
Del manual técnico de las Escalas de Inteligencia Stanford-Binet, quinta edición (SB5): "El uso de un modelo jerárquico de la inteligencia (con un factor global g y factores múltiples en un segundo nivel)... se repite en el SB5." (Roid, 2003b, p. 6)

La interacción histórica entre teorías de la inteligencia y el desarrollo de pruebas específicas ha sido rara. Varias de las pruebas más usadas se elaboraron con sólo una base teórica informal. Una vez creada, estas pruebas estimularon la investigación acerca de sus implicaciones teóricas. Pocas pruebas han tenido derivaciones directas de una teoría particular; sin embargo, las más recientes, tanto revisiones de pruebas más antiguas como pruebas por completo nuevas, parecen tener mayor influencia de las consideraciones teóricas. Haremos notar estas influencias al describir pruebas específicas en los capítulos 8 y 9.

Dos teorías clásicas

Hay dos teorías clásicas de la inteligencia que dominaron la literatura sobre la naturaleza de este concepto. Empezamos la descripción de las teorías con estas dos aproximaciones.

La "*g*" de Spearman «178-179a

Charles **Spearman** (1904, 1927a, 1927b), de origen inglés, desarrolló lo que se suele considerar la primera teoría basada en datos acerca de la capacidad mental humana.

En el capítulo 1 se presentó una breve descripción de esta teoría, para la cual Spearman se basó en el examen de las correlaciones entre muchas pruebas de funciones sensoriales simples. Pensó que estas correlaciones eran lo suficientemente altas para concluir que el desempeño en las pruebas dependía, en su mayor parte, de una capacidad mental general, a la que llamó capacidad general **"*g*"** (siempre con minúscula). Desde luego, las correlaciones entre las diversas medidas no eran perfectas, sino que cada prueba tenía cierta varianza única o específica independiente de "g". Así, cualquier conjunto de pruebas tenía una serie de factores "s" y un factor "g". Spearman también relegó la varianza de error a los factores "s". De ahí que cada uno de ellos contenía cierta varianza única correspondiente a una capacidad específica más una varianza de error. Sin embargo, muchos resúmenes de la teoría de Spearman mencionan sólo "g" y "s".

La figura 7-2 describe la teoría de Spearman. Cada óvalo de la figura representa una prueba. El grado en que se superponen entre sí los óvalos representa el grado de correlación entre ellos. El área grande en el centro corresponde a "g", el factor general de la capacidad mental. Cada óvalo también tiene un área que no se superpone a ningún otro óvalo; estas áreas donde no hay superposición son los factores "s", específicos de una prueba particular.

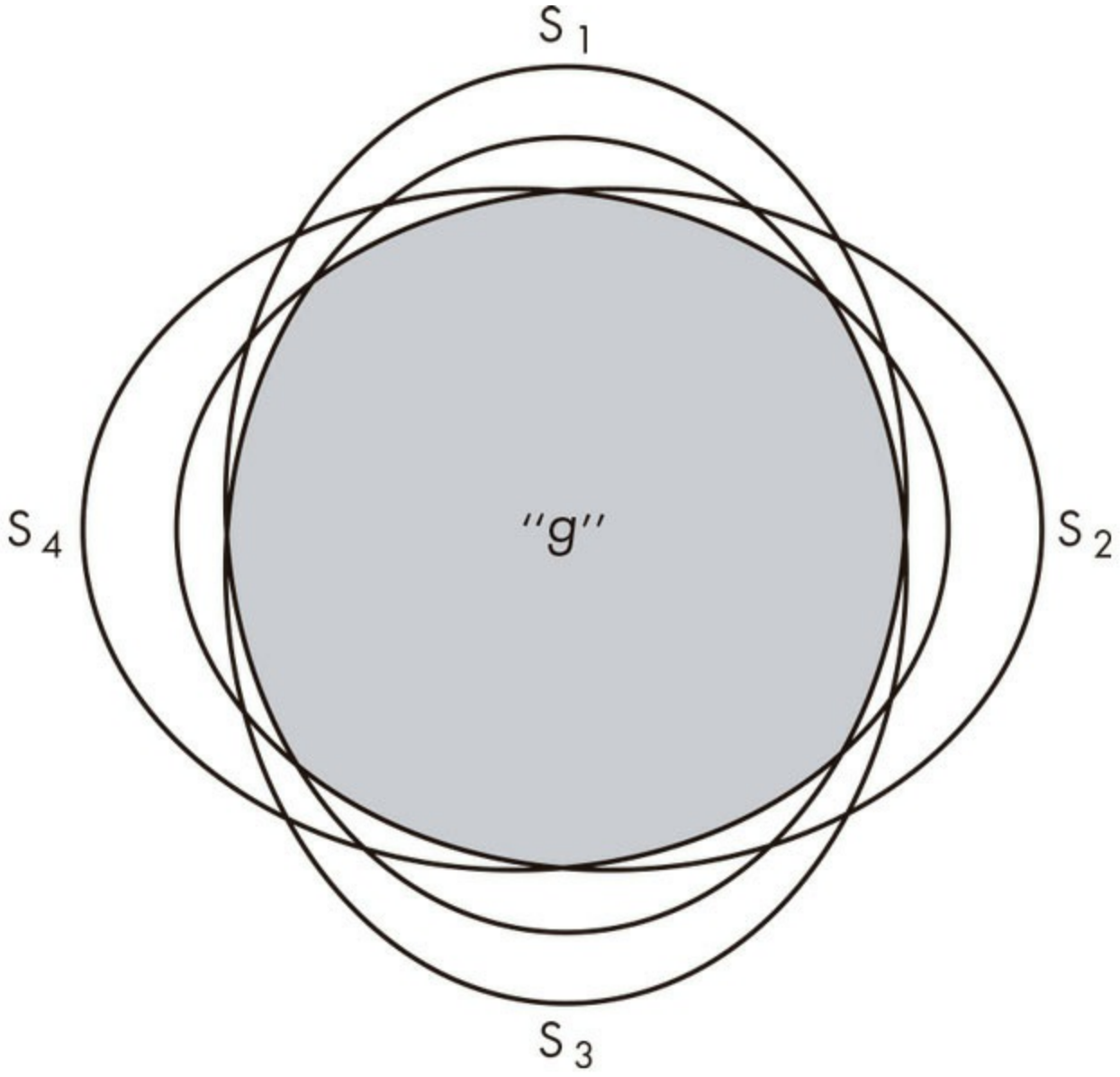


Figura 7-2. Ilustración de la teoría de la "*g*" de Spearman.

Ya que la teoría tiene dos tipos de factores ("g" y una serie de "s"), Spearman la llamó **teoría bifactorial**. Sin embargo, el factor dominante en la teoría es "g", mientras que los factores "s" no son de mucho interés. De ahí que, a pesar de que Spearman usa la terminología bifactorial, su teoría de la inteligencia a menudo se denomina de un factor o unifactorial. A veces sólo se le llama teoría de "g".

Resumen de puntos clave 7-2

Principales teorías de la inteligencia

- Dos teorías clásicas
- La "g" de Spearman
- Capacidades mentales primarias de Thurstone
- Modelos jerárquicos
- Teorías del desarrollo
- Modelos del procesamiento de información y biológicos

En el desarrollo de su teoría de la inteligencia humana, Spearman trabajó los elementos de análisis factorial. Ya antes revisamos esta técnica estadística (127a»). Para los estándares actuales, sus métodos eran bastante primitivos; sin embargo, él mostró una notable intuición respecto de cómo pensar en las relaciones entre varias pruebas, con lo que marcó el camino para una amplia variedad de aplicaciones en las pruebas y en las ciencias sociales.

El factor "g" de Spearman sigue siendo un concepto central en el pensamiento psicológico acerca de la inteligencia, pues sirve como punto de referencia común en los manuales de las pruebas, así como en otras teorías de la inteligencia. Debemos notar que los trabajos originales de Spearman, aunque ahora son anticuados en muchos aspectos, constituyen una fuente rica en ideas para los estudiantes contemporáneos de psicología. Muchos resúmenes contemporáneos sobresimplifican su pensamiento. Por ejemplo, además del factor "g", creó los factores "w" (will [voluntad]) y "c" (carácter), a los que atribuyó efectos en el desempeño en la prueba. Sin embargo, "g" es el concepto central cuya influencia más ha perdurado en el campo.

Capacidades mentales primarias de Thurstone

A lo largo de los primeros años del debate sobre la naturaleza de la inteligencia, el psicólogo estadounidense **L. L. Thurstone**, de la Universidad de Chicago, creó la principal competidora de la teoría de Spearman de "g". Mientras que Spearman decía que las correlaciones entre diferentes pruebas eran lo suficientemente altas para pensar que, en su mayor parte, medían un factor común, Thurstone (1938) creía que las correlaciones eran lo suficientemente bajas para pensar que medían diversos factores, en gran parte independientes, lo que dio por resultado la **teoría multifactorial**. La figura 7-3 representa la teoría de Thurstone. Igual que en la ilustración de la teoría de Spearman, el grado de superposición entre los óvalos representa el nivel de correlación. Thurstone hizo hincapié en la separación entre los óvalos, mientras que Spearman lo hizo en la superposición. Cada una de las "P" de la figura 7-2 es un factor relativamente independiente. Como Spearman al desarrollar su teoría, Thurstone hizo contribuciones importantes a la metodología analítico-factorial. Sus libros, *The Vectors of the Mind* [Los vectores de la mente] (Thurstone, 1938), y sobre todo su revisión, *Multiple-Factor Analysis* [Análisis factorial múltiple] (Thurstone, 1947), ayudaron a definir el análisis factorial moderno.

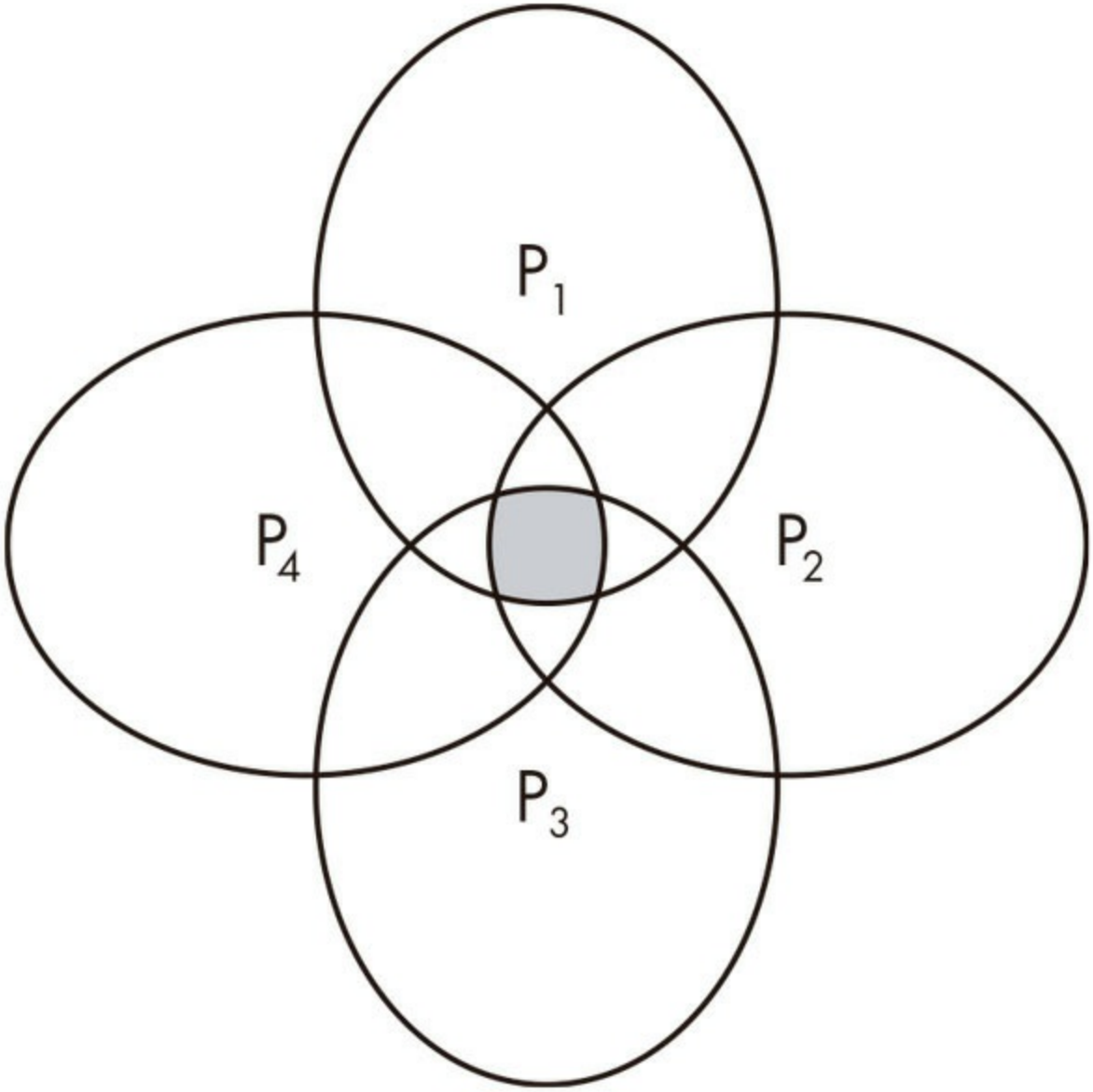


Figura 7-3. Ilustración de la teoría de Thurstone de las capacidades mentales primarias.

En su estudio más famoso, Thurstone (1938) aplicó una batería de 60 pruebas (¡que duraron 15 horas!) a 240 estudiantes. (La muestra fue sumamente selectiva: todos eran hombres y estudiantes de la Universidad de Chicago.) Thurstone extrajo 12 factores, nueve de los cuales consideró interpretables; los llamó factores grupales o **capacidades mentales primarias**, que fue el término que se consolidó. En el cuadro 7-2 aparece una lista con los nueve factores originales que Thurstone identificó y una breve descripción de cada uno.

Cuadro 7-2. Las nueve capacidades mentales primarias originales de Thurstone

S	Espacial	Capacidad espacial, especialmente visual, como al hacer rotaciones mentales de figuras geométricas o al contar cubos que están escondidos
P	Perceptual	Percepción, especialmente velocidad de percepción visual, como al examinar una página impresa para identificar letras o al comparar columnas de números
N	Numérica	Velocidad y exactitud especialmente numéricas para el cálculo
V	Verbal	Capacidad verbal, que incluye analogías, antónimos, comprensión de lectura
M	Memoria	Memoria a corto plazo, como en el aprendizaje de pares asociados
W	Palabras	Fluidez con las palabras, especialmente al tratar con palabras aisladas, como en una prueba de palabras desordenadas o de fluidez con las palabras
I	Inducción	Encontrar una regla o principio que resuelva un problema, como en series de números, clasificación de figuras o analogías de patrones
R	Razonamiento	Razonamiento, especialmente al tratar con problemas de soluciones predeterminadas, como en el razonamiento aritmético
D	Deducción	Factor tímidamente definido por diversas pruebas que apelan a la aplicación de una regla

Es interesante que Thurstone fuera el único teórico importante que también fue autor de pruebas de capacidad mental que alcanzaron un uso generalizado. Hubo ediciones de distintas editoriales y para distintos niveles de edad, pero en la actualidad ninguna de ellas se usa. Es comprensible que todas las pruebas incluyeran la etiqueta de "capacidades mentales primarias" (CMP), que nosotros emplearemos aquí como un descriptor genérico. Las diversas versiones de CMP cubrían sólo cinco de los nueve factores originales, pero no siempre eran los mismos cinco. Así, hay muchas referencias en la literatura a los cinco factores de Thurstone, pero es fácil terminar confundido tratando de identificar cuáles son exactamente los cinco factores en cada caso. Como se resume en el cuadro 7-3, de los nueve factores originales, cuatro aparecen en casi todas las versiones de las pruebas de CMP: espacial, numérica, verbal y razonamiento. Los factores originales de inducción, razonamiento y deducción se fundieron en el de razonamiento, mientras que los factores perceptual, memoria y fluidez con las palabras aparecen en algunas pruebas de CMP, pero en otras no, de modo que siempre incluyen en total cinco factores. De estos tres últimos factores, el perceptual era casi siempre el quinto factor de una prueba de CMP.

Cuadro 7-3. Pruebas que aparecen en distintas versiones de las pruebas de capacidades mentales primarias (CMP)

Factores originales	**En la mayoría de las pruebas de CMP**	**En algunas pruebas de CMP**
Espacial	X	
Numérica	X	
Verbal	X	
Inducción		
Razonamiento	X	

Deducción		
Perceptual		X
Memoria		X
Fluidez con las palabras		X

Thurstone no fue el único que propuso una teoría multifactorial de la inteligencia. En lo que llamó la *estructura del modelo del intelecto*, J. P. Guilford (1956, 1959b, 1967, 1985, 1988) presentó lo que es, sin duda, la versión más extrema de una teoría multifactorial de la inteligencia. De acuerdo con Guilford, la capacidad mental se manifiesta a lo largo de tres ejes principales: contenidos, productos y operaciones. Cada uno de estos ejes contiene sus propias subdivisiones: cinco de contenido, seis de productos y seis de operaciones. Los tres ejes pueden representarse en forma de cubo con subdivisiones que forman celdas, de modo que resultan $5 \times 6 \times 6 = 180$ celdas, las cuales Guilford postuló como relativamente independientes entre sí.[2]

La teoría de Guilford no resistió la prueba del tiempo (y de la investigación), pero una de las distinciones que conformaron el modelo ha permanecido, es decir, la distinción entre producción divergente y producción convergente. Éstas eran subdivisiones a lo largo del eje de las operaciones. La **producción divergente** implica producir soluciones alternas o inusuales, mientras que la **producción convergente** implica identificar una sola respuesta correcta. Es decir, en el pensamiento convergente, la mente converge en una respuesta, mientras que en el divergente, la mente se desvía del camino usual para encontrar diversas posibilidades. Esta referencia al pensamiento divergente ayudó a estimular una gran cantidad de investigaciones sobre el pensamiento creativo.

¡Inténtalo!

Responde estas preguntas, que ilustran la diferencia entre pensamiento convergente y divergente.
Convergente: ¿Cuál es el uso más común de un ladrillo? (Una respuesta correcta.)
Divergente: ¿De cuántos modos diferentes puedes usar un ladrillo? (Muchas respuestas posibles.)

Modelos jerárquicos

El debate "uno contra muchos", que protagonizaron Spearman y Thurstone, ha demostrado ser una de las batallas más duraderas de la psicología. Los **modelos jerárquicos** de la inteligencia adoptan una posición de compromiso, pues admiten que hay muchas capacidades separadas, pero notan que están organizadas en una jerarquía en la que sólo uno o dos factores dominantes están en la parte superior. Se han propuesto varios modelos jerárquicos, de los cuales examinaremos tres.

Primero, agreguemos esto al margen. Como señalamos antes, el desarrollo de las teorías de Spearman y Thurstone fue de la mano con los desarrollos en el análisis factorial. Así fue, también, con los modelos jerárquicos. De especial importancia para

estas teorías fueron las nociones de rotación oblicua (como opuesto a ortogonal) de los ejes, factores de segundo orden (e, incluso, de más alto orden) y, en fechas más recientes, análisis factorial confirmatorio y modelo de ecuaciones estructurales. Explicar estos temas más avanzados nos llevaría demasiado lejos para lo que pretendemos en este libro introductorio; sin embargo, el lector debe estar enterado de que los modelos jerárquicos dependen de estas metodologías.

La figura 7-4 presenta una versión generalizada de los modelos jerárquicos. Los modelos específicos varían en sus detalles, pero todos siguen este patrón general. Podemos notar las siguientes características: en los niveles inferiores (en la parte baja de la figura), encontramos capacidades muy específicas, algunas de las cuales tienen correlaciones más altas entre sí, de modo que forman grupos, que se representan como factores secundarios. Siguiendo hacia arriba en esta figura, algunos de los factores secundarios tienen correlaciones más altas, de modo que conforman los factores principales. Y entonces, éstos muestran cierto grado de correlación y se resumen en "g", es decir, la capacidad mental general.

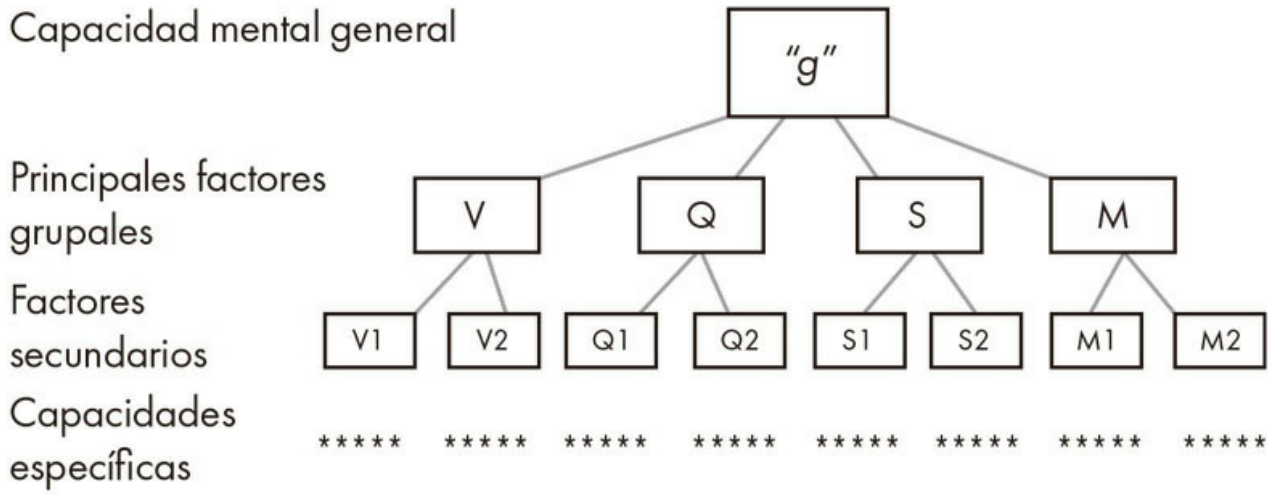


Figura 7-4. Versión generalizada de un modelo jerárquico de la inteligencia.

Inteligencia fluida y cristalizada de Cattell «182a

Más o menos contemporáneo del trabajo de Guilford, R. B. Cattell[3] entró en la lucha teórica con su teoría de la inteligencia fluida y cristalizada. En el primer trabajo de Cattell (1940) sobre este tema, critica mordazmente a Binet por producir una prueba demasiado verbal y dependiente de la escolarización. Propuso una "prueba de inteligencia perceptual" basada en su mayor parte en reactivos hechos de figuras, como matrices y laberintos, por lo que la llamó prueba neutral en términos culturales. Después, Cattell (1963) elaboró y refinó (Horn & Cattell, 1966) la teoría Gf-Gc. Gc, una **inteligencia cristalizada** general, es la suma de todo lo que una persona ha aprendido: un fondo de

información, relaciones y habilidades mentales desarrolladas a lo largo de la educación, la experiencia y la práctica. Gf, una **inteligencia fluida** general, podría considerarse como la fuerza mental bruta, y es probable que tenga un sustrato neurológico. La diferencia entre Gf y Gc corresponde, aunque no exactamente, a la diferencia entre las influencias hereditarias y ambientales sobre la inteligencia. La diferencia podría pensarse también como la que existe entre los términos "potencial" y "real" (¡conceptos que vienen directamente de Aristóteles!).

Tanto Gf como Gc están compuestos de varios factores más específicos, es decir, existen varios componentes de Gf y varios de Gc. Así, la teoría se considera como modelo jerárquico. Hay algunas diferencias de opinión respecto de si Gf y Gc, en última instancia, se mezclan en una especie de super "g". En Hunt (2011) y Sternberg y Kaufman (2011) se presentan resúmenes recientes de esta teoría.

Cattell realizó gran parte de su trabajo sobre esta teoría en colaboración con J. L. Horn, por lo que a veces se le llama teoría de Cattell-Horn. Más a menudo se conoce como teoría de la inteligencia fluida y cristalizada de Cattell, o simplemente teoría Gf-Gc. Como sea que se le llame, la teoría ha demostrado tener un gran atractivo para los psicólogos. Gf es de especial interés; ¿podemos medir esta fuerza mental bruta, ajena a las influencias culturales, la educación, los antecedentes en el hogar? ¿Pero cómo se podría acceder a Gf si no es mediante la manifestación de las capacidades desarrolladas, que son Gc por definición? Algunos modelos de procesamiento de información que se consideran en las siguientes secciones intentan abordar justo esta cuestión.

Modelo de Vernon

Philip Vernon (1947, 1950, 1961, 1965) elaboró uno de los primeros modelos jerárquicos de la inteligencia. A diferencia de la mayoría de los teóricos que citamos aquí, Vernon mismo hizo poca investigación original; trató de resumir de manera práctica la enorme cantidad de investigaciones realizadas por otros hasta cerca de 1950 y de darle cierta unidad a las orientaciones teóricas en conflicto. Su primer resumen se publicó en 1950; fue ligeramente actualizado en 1961 y elaborado nuevamente en 1965.

De acuerdo con el modelo de Vernon, existe una serie de capacidades definidas de manera restringida, las cuales tienden a agruparse en varios "factores grupales secundarios". (Decir que algunas capacidades específicas "se agrupan" significa que algunas tienen correlaciones más altas entre sí.) Entonces, los factores grupales secundarios se agrupan en dos categorías principales o factores. Vernon los llamó v:ed (verbal: educativo) y k:m (espacial: mecánico). A veces se denomina factor "práctico" al factor k:m. Los dos principales factores grupales están relacionados entre sí y forman una capacidad mental general, global. Ésta es la "g" de Spearman.

Vernon no intentó especificar de manera exacta cuántos factores grupales secundarios existen, pero mencionó, al menos, qué son estos factores. Destacan las capacidades verbal y numérica entre los factores grupales secundarios de v:ed en el modelo de Vernon. Los factores secundarios de k:m incluyen capacidad espacial, información

mecánica y capacidades psicomotrices.

Resumen de Carroll

John Carroll trabajó en la viña del análisis factorial durante muchos años. En la década de 1980, empezó a resumir cientos de análisis factoriales acerca de las capacidades humanas. Su obra monumental, *Human Cognitive Abilities: A Survey of Factor Analytic Studies* [Capacidades cognitivas humanas: una investigación de los estudios analítico-factoriales] (Carroll, 1993), concluye con su propio resumen de un modelo jerárquico. La figura 7-5 presenta el modelo de Carroll.

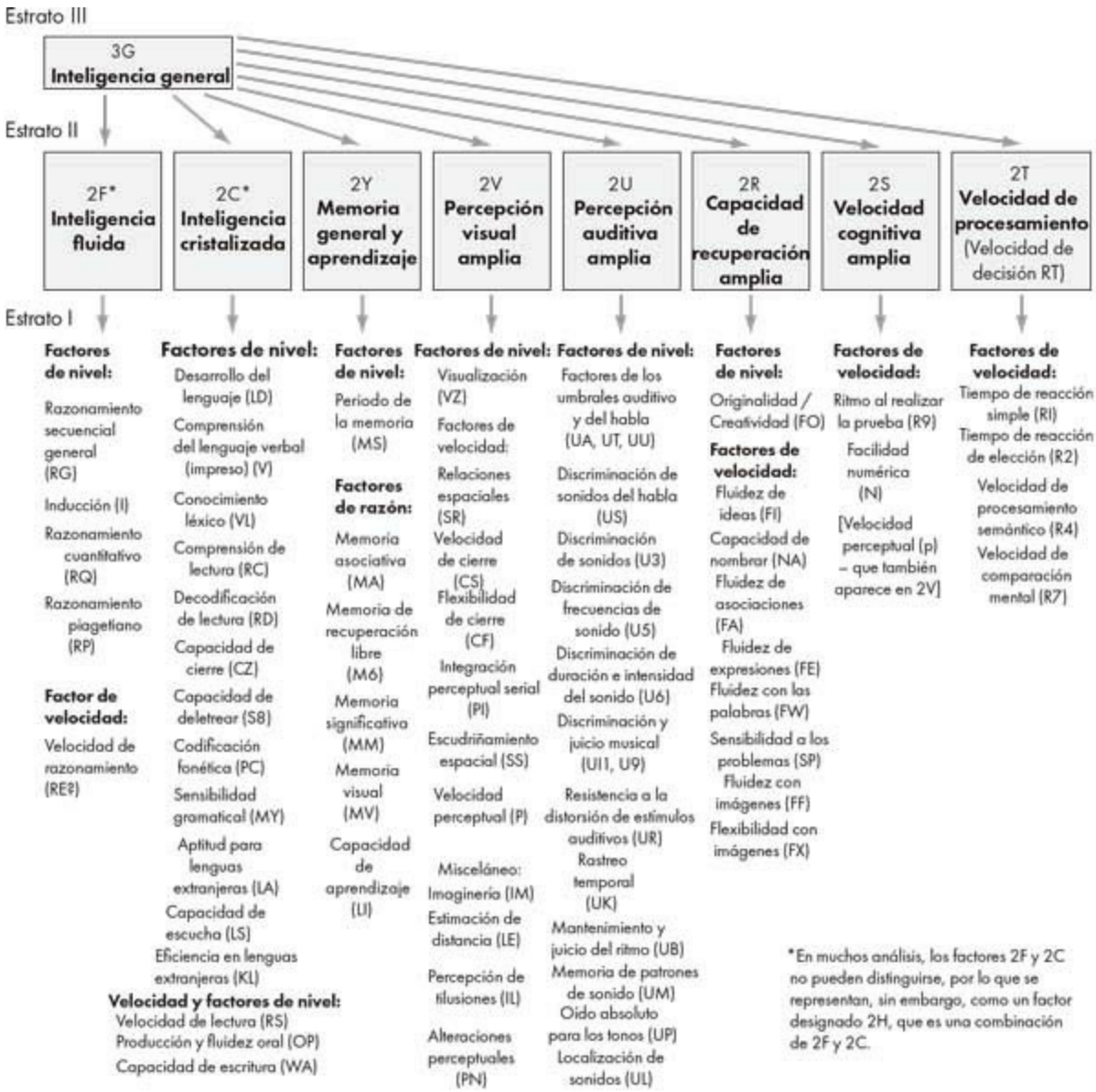


Figura 7-5. Modelo jerárquico de tres estratos de la inteligencia humana de Carroll. «184a «184b

Fuente: Carroll, J. B. Human cognitive abilities: A survey of factor analytic studies (p. 626). Copyright © 1993.
Reproducida con autorización de Cambridge University Press.

Carroll usa tres estratos; de hecho, él le llama **teoría de tres estratos**. La inteligencia general está en el nivel más alto; es la "g" de Spearman. Carroll incorpora Gf y Gc de Cattell en el segundo nivel; sin embargo, hay varios factores grupales de segundo nivel además de Gf y Gc, algunos de los cuales corresponden a las capacidades mentales primarias de Thurston (véase cuadro 7-2). Por último, en el estrato I, hay una gran cantidad de capacidades más específicas y de definición restringida. Carroll nota que algunas de estas capacidades específicas son complejas en términos factoriales, por lo que contribuyen a más de un factor grupal. En la figura 7-5, la extensión de las líneas desde la celda de inteligencia general hasta las de los factores grupales indican más o menos la relación entre "g" y un factor grupal: mientras más corta es la línea, mayor es la relación. Por ejemplo, la línea de Gc a g es relativamente corta, lo cual significa que Gc tiene una relación muy alta con el factor g global, aunque Gf está aún más relacionada con g. Hay una línea larga de g a 2R: capacidad de recuperación amplia, porque 2R está menos relacionada con g. El resumen de Carroll que se muestra en la figura 7-5 merece ser estudiado con atención. Probablemente es el mejor resumen actual de todas las aproximaciones analítico-factoriales a la definición de la inteligencia humana. Sternberg y Kaufman (1998) afirmaron que Carroll "integra magistralmente una literatura analítico-factorial extensa y diversa, lo cual le da una gran autoridad a su modelo" (p. 488). Lubinski (2004, p. 98) señaló que el modelo de tres estratos de las capacidades cognitivas de Carroll (1993) es, sin duda, el tratamiento más definitivo del tema".

Teorías del desarrollo

Las teorías clásica y jerárquica que hemos tratado aquí a veces se denominan **teorías psicométricas** de la inteligencia, y dependen principalmente del análisis de las relaciones entre pruebas específicas. Sin embargo, algunas teorías de la inteligencia humana surgen de otras consideraciones o perspectivas, como la del desarrollo. El elemento clave de estas **teorías del desarrollo** es cómo se desarrolla la mente con la edad y la experiencia.

Antes de examinar teorías del desarrollo específicas, notemos que éstas, sean de la inteligencia o de algún otro rasgo, tienden a tener las siguientes características. La figura 7-6 ayuda a resumirlas. Primero, estas teorías **se basan en etapas**; el desarrollo transcurre a través de una serie de etapas, cada una de las cuales tiene características especiales que la hacen cualitativamente diferente de las otras. Por ello, el desarrollo no es sólo la acumulación de más de lo mismo, por ejemplo, de información o vocabulario. Es como la oruga y la mariposa: la mariposa no es sólo una oruga más grande. Ésa es la razón de que el desarrollo, en la figura 7-6, transcurra como una serie de pasos discretos

en vez de una curva típica continua de crecimiento. Segundo, la secuencia de las etapas es invariante; todos pasan por las etapas en el mismo orden, aunque no necesariamente a la misma edad. Así, no es posible saltarse una etapa. Si hay cuatro etapas en la teoría (suponiendo que sea correcta), un niño pasa de la etapa A a la B, de la B a la C y así sucesivamente, pero no puede pasar de manera directa de la etapa A a la D.

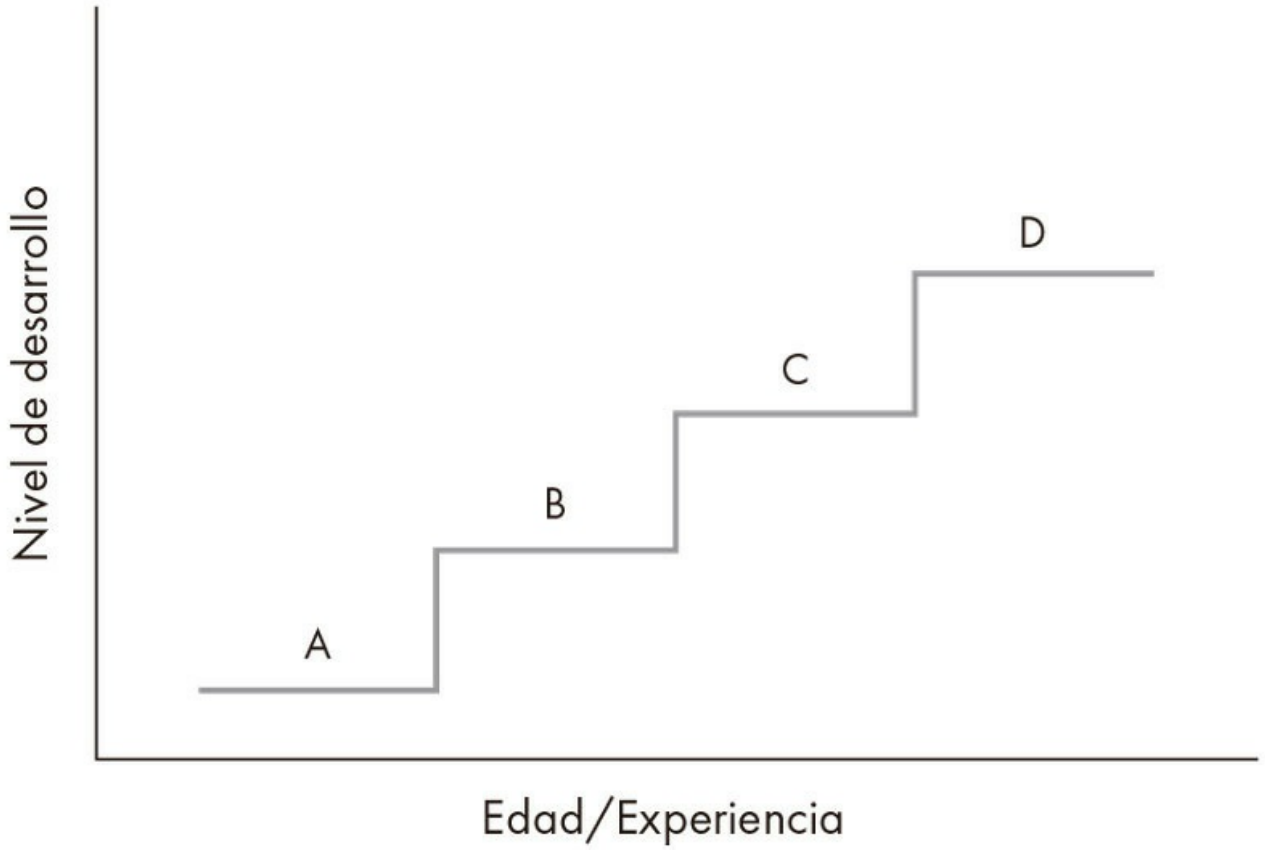


Figura 7-6. Ilustración de las teorías de las etapas.

Tercero, las etapas son irreversibles. Una vez que se alcanza la etapa C, no es posible regresar a la B, así como la oruga se convierte en mariposa, pero la mariposa no puede volverse oruga. Por último, suele haber (aunque no siempre) una relación entre la evolución a través de las etapas y la edad. Por ejemplo, en promedio, los niños pueden alcanzar la etapa C a los 7 años de edad, o la etapa D a los 12, aunque puede haber variabilidad alrededor de estos promedios. Esta evolución por edad puede suponer ciertas experiencias comunes. Un niño que crece en un armario no se desarrollará. Desde luego, estas características están idealizadas; en la práctica, una buena teoría de etapas se aproximará a estas condiciones, pero puede no ajustarse con exactitud a ellas.

Las teorías basadas en etapas se presentan en otras áreas de la psicología además de la inteligencia. Por ejemplo, hay teorías del desarrollo de la personalidad, entre las cuales la teoría del desarrollo psicosocial de Erikson es la más conocida. También hay una teoría

muy conocida de las etapas del dolor.

Teoría del desarrollo cognitivo de Piaget

Sin duda, la teoría del desarrollo de la inteligencia más destacada es la de Jean Piaget, la cual ha tenido una influencia enorme en la psicología del desarrollo y en la educación temprana de niños. De acuerdo con Piaget, la mente humana se desarrolla a través de cuatro etapas (véase, p. ej., Piaget, 1950, 1983; Piaget & Inhelder, 1969), las cuales se bosquejan en el cuadro 7-4.

Cuadro 7-4. Principales etapas del desarrollo intelectual de acuerdo con Piaget

Nombre de la etapa	Edades típicas	Algunas conductas de las etapas
Sensoriomotriz	0 – 2 años	Entrada de estímulos; no hay permanencia del objeto
Preoperacional	2 – 6 años	Uso de palabras para simbolizar; no hay principios de conservación
De operaciones concretas	7 – 12 años	Uso de los principios de conservación y reversibilidad
De operaciones formales	más de 12 años	Pensamiento adulto maduro en términos de hipótesis, causa y efecto

Aunque ha tenido una gran influencia en la psicología y la educación en general, la teoría de Piaget no ha tenido tanta influencia en el campo de las pruebas. La investigación de Piaget tiene su propio cuadro de tareas equivalentes a las pruebas; sin embargo, no se han abierto camino en la clase de pruebas que usan los psicólogos en su trabajo clínico. Esto es más bien curioso, porque las tareas son sencillas y con facilidad se podrían incorporar en las pruebas de inteligencia de aplicación individual que examinamos en el siguiente capítulo. Algunas de ellas incluso podrían adaptarse al formato de opción múltiple que se puede usar en la aplicación grupal. Sin embargo, lo que ocurre es que las tareas piagetianas han quedado en su mayor parte confinadas en la esfera de los programas de investigación de laboratorio.

¡Inténtalo!

¿Puedes identificar alguna diferencia radical en tu modo actual de pensar el mundo y el que tenías, digamos, hace 10 o 15 años, sin que tal diferencia se limite sólo a la cantidad de información? Si puedes identificar esa transformación, esto puede representar parte de una teoría de las etapas de la capacidad mental.

Teorías del procesamiento de información y biológicas de la inteligencia

Las teorías del procesamiento de información y biológicas de la inteligencia humana son, en parte, distintas, pero también tienen varios puntos de encuentro. Un **modelo de procesamiento de información** hace hincapié no en el contenido de lo que se sabe sino en cómo se procesa el contenido. El procesamiento de la computadora a menudo sirve como analogía para los modelos del procesamiento de información. Los **modelos biológicos**, en cambio, hacen hincapié en el funcionamiento cerebral como base para comprender la inteligencia humana. Es decir, cualquier cosa que sea la inteligencia humana, debe funcionar en el cerebro. Así, necesitamos estudiar la actividad cerebral para comprender la inteligencia. Desde luego, las redes neurales del cerebro a menudo se comparan con el procesamiento de la computadora, por lo que los dos tipos de modelos no son por completo distintos uno del otro.

Tareas cognitivas elementales (TCE) «185-187a

En muchas de las aproximaciones del procesamiento de información para comprender la inteligencia humana, un elemento esencial es la **tarea cognitiva elemental** o **TCE**, la cual es una tarea relativamente sencilla que requiere algún tipo de procesamiento mental. Los investigadores esperan que el desempeño en la TCE abra una ventana al funcionamiento mental. De hecho, la TCE puede constituir una medida más o menos directa de la eficiencia del procesamiento mental, algo que está en la raíz de la inteligencia. Ya que las tareas son relativamente sencillas, en apariencia son ajenas a las experiencias educativas y culturales. Por ello, algunos investigadores esperan que el desempeño en las tareas proporcione una medida de inteligencia libre de sesgos culturales.

La literatura de investigación contiene una gran cantidad de TCE, de las cuales presentaremos algunos ejemplos. Una es el tiempo de reacción simple, en la que una persona responde a la aparición de una luz en el centro de una caja, como se muestra en la figura 7-7, bajando un interruptor. Consideremos ahora lo que debe suceder: el individuo debe observar la aparición de la luz, luego debe "decidir" bajar el interruptor y, por último, emitir una respuesta motriz. Aunque esto no parece una conducta inteligente, la velocidad con que una persona puede ejecutar todas estas funciones puede ser el sustrato de la inteligencia humana.

Figura 7-7. Dispositivo para medir los tiempos de reacción.

Ahora consideremos lo que se denomina tiempo de reacción de decisión. Regresemos a la figura 7-7. Una luz puede aparecer del lado izquierdo o derecho; la persona baja el interruptor si la luz aparece del lado derecho, pero no si aparece del lado izquierdo. Consideremos lo que pasa: la persona debe percibir una luz, luego decidir si se trata de la luz meta (la de la derecha) y, por último, ejecutar una acción. La tarea puede hacerse más complicada, y seguir siendo muy sencilla, añadiendo más luces. Otra variación de esta tarea es tener el dedo de la persona presionando el interruptor y luego moverlo hacia la luz cuando ésta aparezca. En esta versión, medimos por separado el tiempo requerido para la respuesta motriz (mover la mano) y otros aspectos de la reacción total.

He aquí otro ejemplo de una TCE. Se presentan dos letras a la persona en un monitor de computadora, como los ejemplos que aparecen en la figura 7-8. La tarea consiste en indicar si las letras son idénticas en términos físicos. Por ejemplo, nuestra tarea puede ser indicar si las letras tienen nombre idéntico, es decir, si son la misma letra (p. ej., aA) aunque no sean idénticas en términos físicos.

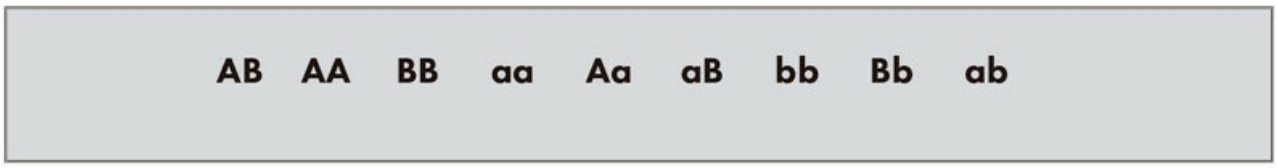


Figura 7-8. Ejemplos de pares de letras usados en TCE.

Una TCE más que utiliza letras es la tarea de verificación semántica. Se emplean tres

letras A, B y C en diferente orden; luego una afirmación en que se relacionan, por ejemplo, “A después de B” o “C entre A y B”. La persona debe indicar si las afirmaciones son verdaderas o falsas. La figura 7-9 muestra algunos ejemplos. Los reactivos aparecen en la pantalla en una rápida sucesión.

Letras en el monitor	Afirmación	Marca de verdadero (V) o falso (F)
B C A	**B entre C y A**	**V F**
C B	**B después de C**	**V F**
A C B	**A antes de C y B**	**V F**

Figura 7-9. Ejemplos de reactivos en una tarea de verificación semántica.

Un último ejemplo es el tiempo de inspección. En esta tarea, el individuo ve dos líneas paralelas que se iluminan con destellos de luz en un taquistoscopio o monitor de computadora. La tarea consiste nada más en decir cuál es la línea más grande. Otra vez vemos el ahora conocido paradigma: entrada de información sensorial, codificación, determinación, comparación, reacción. Stokes y Bohrs (2001) ofrecen un ejemplo de una tarea de tiempo de inspección en la que se emplean letras. En Deary y Stough (1996) y Nettelbeck (2011) se puede encontrar una descripción más detallada de estas tareas y su relación con la inteligencia.

En todos estos ejemplos, el investigador puede medir en varios ensayos no sólo el tiempo promedio de reacción, sino también la variabilidad (desviación estándar) del tiempo de reacción. Desde luego, con tiempo suficiente –y no se necesita mucho– casi cualquiera obtendría puntuaciones perfectas en estas tareas, pero ¿qué tan bueno es el desempeño cuando los reactivos aparecen con rapidez? Es posible contrastar puntuaciones de diferentes versiones de la tarea, por ejemplo, identidad física frente a versiones de la misma letra.

La **memoria de trabajo** ha surgido del enfoque del procesamiento de información como una posible base importante de la inteligencia o, al menos, uno de sus componentes (véase Barrouillet & Gaillard, 2011; Conway, Getz, Macnamara, & Engel de Abreu, 2011; Cowan, 2005; Hunt, 2011; Nettlebeck, 2011). A riesgo de simplificar demasiado, pensémoslo de este modo: retener en la memoria varios hechos y conceptos. Qué tanto podemos manipular hechos y conceptos, relacionarlos entre sí y hacer algo con ellos, por ejemplo, tomar una decisión, pero con rapidez y, además, mientras se sostiene una conversación no relacionada con esto. La investigación de la memoria de trabajo trata de hacer un modelo de cómo ocurre todo eso, con qué grado de eficiencia se lleva a cabo y cómo se podría relacionar con la inteligencia. ¿Es ésta la esencia de “g” fluida?

Hemos presentado varios ejemplos de los tipos de tareas que se usan en el enfoque del procesamiento de información para estudiar la inteligencia. Lo esencial es que, en todas estas tareas, los investigadores esperan que el procesamiento de información nos diga algo acerca de la naturaleza básica de la inteligencia libre de influencias culturales. Esto es

lo que Brody (1992) llamó “la búsqueda del santo grial” (p. 50). Pero Hunt (2011), invocando la misma leyenda de Arturo, señaló que “el éxito de la búsqueda de una sola función de procesamiento de información que explique la inteligencia ya no es más probable que el éxito en la búsqueda del santo grial” (p. 160).

Teoría de Jensen

Uno de los principales defensores del enfoque del procesamiento de información a la inteligencia es Arthur **Jensen**, quien debe su fama (o infamia) a su artículo en la *Harvard Educational Review* sobre las diferencias de la inteligencia en blancos y negros (Jensen, 1969). En la mayor parte de su investigación, publicada en numerosos artículos, se concentró en las relaciones entre las TCE y la inteligencia general. En *The g Factor: The Science of Mental Ability,* Jensen (1998) presentó un resumen integral de su investigación y su posición teórica.

La figura 7-10 resume el modelo de Jensen. La dirección de las flechas indica causalidad en esta figura; en la parte inferior están las tareas de tiempo de reacción, una colección de TCE. Varios procesos de información (P1, P2, etc.) determinan el desempeño en estas tareas. En cambio, las P se determinan por medio de un factor de procesamiento de información general (PI), así como por el puro factor de velocidad del tiempo de reacción (TR). La inteligencia fluida determina el factor PI, pero no el TR. En la otra dirección, la inteligencia fluida es la que determina el desempeño en las pruebas psicométricas, como las que estudiamos en los capítulos 8 y 9.

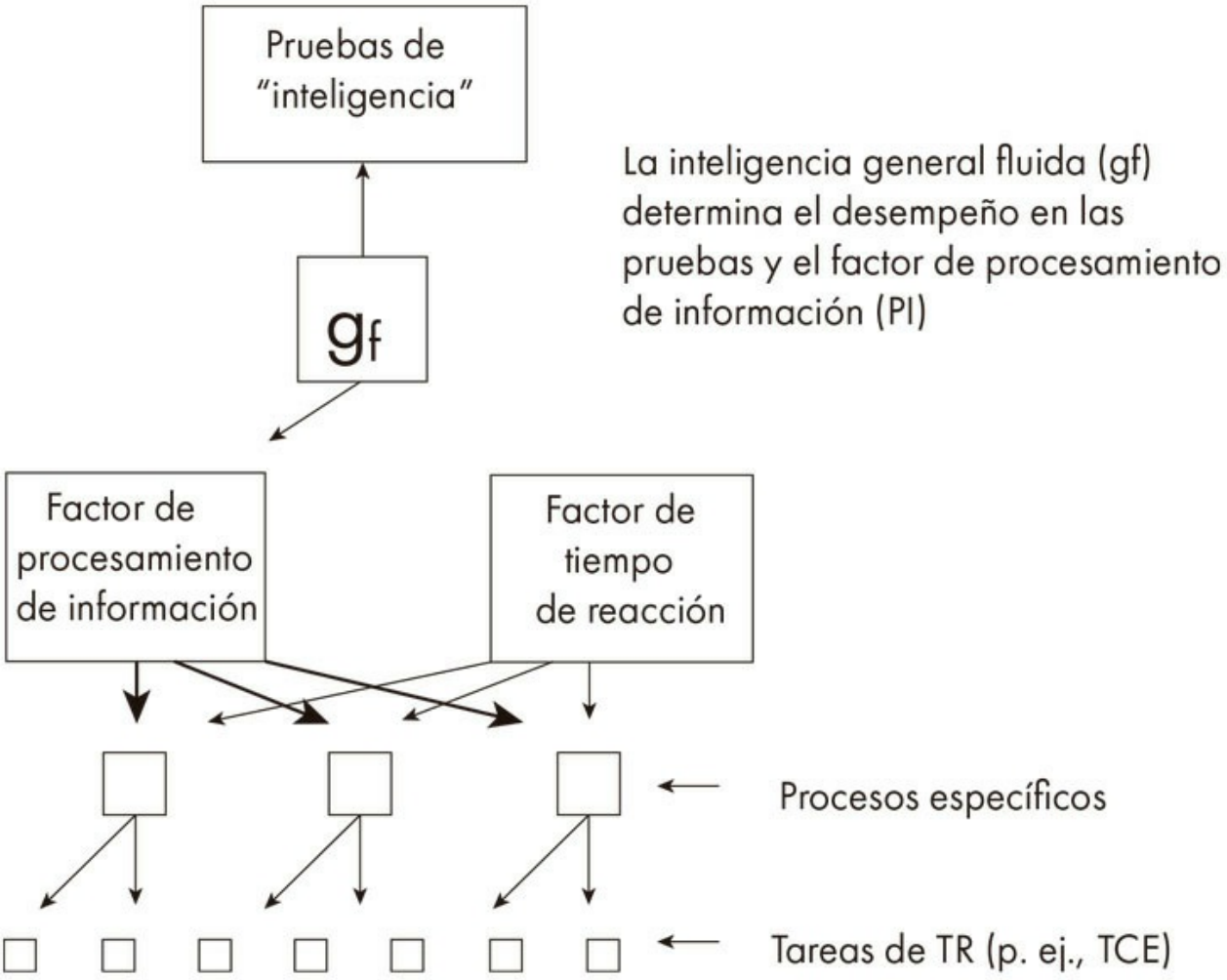


Figura 7-10. Ilustración esquemática del modelo del procesamiento de información de la inteligencia de Jensen.
Fuente: De acuerdo con Jensen, 1998.

Cuál es el principal interés de "g", específicamente de la inteligencia fluida: que "g" determinará el desempeño en las pruebas. Sin embargo, el desempeño en las tareas de TR, de acuerdo con el razonamiento de Jensen, también refleja "g" mediada por ciertos procesos intermedios. En el modelo de Jensen, el desempeño en el TCE proporciona la principal vía hacia "g".

Otras tres teorías

Mencionaremos de manera breve otras tres teorías que se ubican en la categoría de los modelos del procesamiento de información y biológicos. Han aparecido en muchas publicaciones, pero no han tenido mucha influencia en cuestiones prácticas de las pruebas psicológicas. Por eso nuestro tratamiento de ellas es breve.

Teoría triárquica de Sternberg. En una prolífica serie de libros, artículos y presentaciones, Sternberg propuso su **teoría triárquica** de la inteligencia. En Sternberg (2011) se puede consultar un útil y actual resumen de la teoría y el frenesí de sus primeras publicaciones. De acuerdo con esta teoría, la inteligencia tiene tres facetas, cada una de las cuales, a su vez, tiene varias subdivisiones. Una gran cantidad de tríadas pueblan la teoría, la cual tiene tres subteorías: componencial, experiencial y contextual. Ésta es la triarquía en el título de la teoría. La subteoría componencial se refiere a los procesos mentales; el segundo proceso componencial es el desempeño, que implica la solución real de un problema bajo la supervisión de los metacomponentes. El tercer proceso mental especificado en la subteoría componencial es la adquisición de conocimiento.

La subteoría componencial ha sido la parte más citada de la teoría triárquica. Ya que esta parte de la teoría se concentra en los procesos, la clasificamos dentro de las teorías del procesamiento de información. Como resultado adicional de la teoría triárquica, Sternberg a veces hace hincapié en lo que llama "conocimiento tácito" (Sternberg & Wagner, 1986) e "inteligencia práctica", una noción muy similar a la de funcionamiento adaptativo que consideramos en el capítulo 8.

Sternberg ha intentado construir algunas pruebas consistentes con su teoría, pero sin mucho éxito en la práctica. En Gottfredson (2003) se puede leer una crítica fulminante a la noción de inteligencia práctica de Sternberg.

Teoría de las inteligencias múltiples de Gardner. En una auténtica inundación de publicaciones, Howard Gardner propuso su teoría de las **inteligencias múltiples** o **teoría IM**. Clasificamos la teoría de Gardner en la categoría biológica, porque con frecuencia se refiere al funcionamiento cerebral y a conceptos evolutivos en sus escritos, aunque también usa otros criterios. Gardner (1983; véase también Gardner, 1986) anunció primero siete inteligencias: lingüística, musical, lógico-matemática, espacial, corporal-cinestésica, intrapersonal e interpersonal. Las inteligencias lingüística, lógico-matemática y espacial están bien representadas en otras teorías, pero las inteligencias musical, corporal-cinestésica, intrapersonal e interpersonal son particulares de la teoría de Gardner. La mayoría de las personas considerarían estas funciones como algo distinto de la inteligencia, por ejemplo, como parte de un dominio psicomotor o como parte de la personalidad.

En fechas más recientes, Gardner (1999) anunció la adición de tres y, tal vez, cuatro tipos de inteligencia, que incluyen las inteligencias naturalista, espiritual, existencial y moral. Por ejemplo, Gardner define la inteligencia naturalista como "maestría en el reconocimiento y clasificación de numerosas especies –de flora y fauna– del ambiente" (p. 48). Y de nuevo Gardner (2006) apareció con cinco "clases de mentes" más: disciplinada, sintetizadora, creativa, respetuosa y ética. Y la lista sigue creciendo.

La teoría IM de Gardner ha sido muy popular en los círculos educativos; en algunos casos, la teoría ha dado origen a programas de estudio escolares enteros. En términos de las implicaciones educativas, el principal impulso de la teoría de Gardner parece ser doble. Primero, maximizar el potencial de las personas y, segundo, todos somos buenos

en algo. Éstas pueden ser máximas útiles para los enfoques educativos; sin embargo, no constituyen una adecuada teoría de la inteligencia.

Teoría PASS. La **teoría PASS**, elaborada por Das, Naglieri y Kirby (1994), es un modelo de procesamiento de información con referencias explícitas a fundamentos biológicos, en especial a varias áreas del cerebro. Los elementos esenciales de la teoría tienen su origen en el trabajo del neuropsicólogo ruso A. R. Luria sobre el retraso mental y el daño cerebral. La figura 7-11 presenta un bosquejo simplificado de la teoría PASS. Ésta postula tres unidades funcionales, cada una de las cuales ocurre en áreas específicas del cerebro. La primera unidad es la atención y la excitación; se refiere simplemente a que la persona tiene que estar despierta y atenta para captar la información en su sistema. La segunda unidad funcional recibe y procesa la información. Una característica clave de la teoría es que hay dos tipos de procesos: secuencial y simultáneo. El **procesamiento simultáneo** trata con el material holístico, integrado; es decir, se trata todo el material a la vez. El **procesamiento secuencial** se ocupa de los eventos ordenados o encadenados en el tiempo. La tercera unidad funcional implica planear. Lo más común es que esta unidad incluya lo que otras teorías llaman funciones ejecutivas: monitoreo, evaluación, dirección. Das et al. (1994) sugieren que "planear es la esencia de la inteligencia humana" (p. 17).

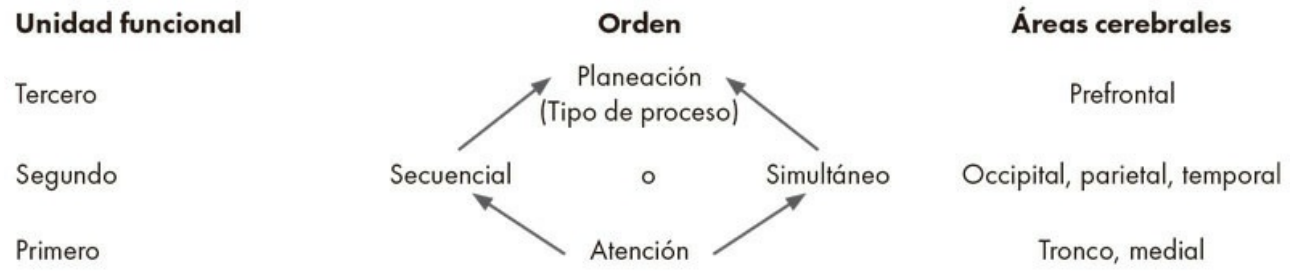


Figura 7-11. Bosquejo del modelo PASS.

Dos pruebas usan explícitamente la teoría PASS en su elaboración e interpretación: el *Kaufman Assessment Battery for Children* [Batería de Evaluación Kaufman para Niños], segunda edición (K-ABC II; Kaufman & Kaufman, 2004) y el *Cognitive Assessment System* [Sistema de Evaluación Cognitiva] (CAS; Naglieri & Das, 1997).

Estatus actual de las pruebas en relación con las teorías

En la práctica contemporánea, notamos las siguientes generalizaciones respecto de la relación entre teorías de la inteligencia y la práctica cotidiana de la evaluación de la capacidad mental.

1. Es claro que, entre las pruebas más usadas, predomina una versión del modelo jerárquico. En la batalla clásica entre "uno" (Spearman) y "muchos" (Thurstone), ambos son ganadores (o perdedores, dependiendo del punto de vista). En la actualidad, todas las pruebas importantes proporcionan una puntuación total, lo cual suele reconocerse como un indicador de "g"; además, proporcionan varias subpuntuaciones que corresponden a factores grupales amplios. Por lo común, estos factores son el verbal, no verbal, espacial, memoria y numérico o cuantitativo. El modelo de Vernon, la teoría Gf-Gc y la teoría de tres estratos de Carroll se citan ampliamente en los manuales de pruebas de capacidad mental contemporáneas. Así, el psicólogo debe tener cierto conocimiento de los modelos jerárquicos para darle sentido a las pruebas de capacidad mental actuales.

2. Los modelos del procesamiento de información y biológicos, hasta la fecha, no han tenido una gran influencia en las pruebas, lo cual es sorprendente. La discusión y la investigación de estos modelos dominan la literatura actual sobre la inteligencia; abundan las obras sobre las tareas cognitivas elementales y la memoria de trabajo en las revistas. Sin embargo, aún no vemos muchos efectos prácticos de estos modelos en la evaluación cotidiana que llevan a cabo los psicólogos. Si están por venir cambios importantes en la medición de la inteligencia humana, es probable que provengan de los modelos del procesamiento de información. Empezamos a ver algunas influencias de la investigación sobre la memoria de trabajo en las pruebas de inteligencia de aplicación individual. El hecho es que el desempeño en las tareas de procesamiento no tiene correlaciones muy altas con las puntuaciones totales de las medidas establecidas de inteligencia, en especial el componente verbal (el cual tiende a ser el que mejor predice en cuestiones prácticas como el éxito escolar o el desempeño en el trabajo). Muchas de las tareas de procesamiento son más bien torpes para su uso ordinario.

3. Las teorías del desarrollo de la inteligencia no han tenido gran influencia en la medición práctica de la capacidad mental, lo cual también es sorprendente. La obra de Piaget ha dominado el pensamiento acerca de la capacidad mental en la psicología del desarrollo por décadas. Sin embargo, ha tenido poco efecto en las pruebas. Las tareas piagetianas siguen quedando en gran medida confinadas a los laboratorios.

Diferencias grupales en la inteligencia

Pocos temas en psicología generan más curiosidad y controversia que las diferencias grupales en la inteligencia. La gente de las culturas occidentales tiene una particular fascinación con estas diferencias, quizá más de lo que sería saludable, pues linda con la obsesión. Nuestra revisión de estas diferencias debería empezar considerando las tres perspectivas, las cuales ayudan a esparcir algo de la carga emocional a menudo asociada con estas diferencias. En realidad, estas perspectivas son importantes independientemente de que tomemos en cuenta las diferencias en la inteligencia o en cualquier otro rasgo, pues se pueden aplicar a las diferencias en la personalidad y en otros rasgos, así como a las diferencias en la inteligencia.

Primero, u*na* superposición en las distribuciones es la regla. Los informes de investigación suelen comunicar las diferencias grupales en forma de promedios. Consideremos el rasgo X, que podría ser inteligencia, introversión o estatura. Las afirmaciones típicas incluyen éstas: el grupo A es significativamente superior al grupo B en X, hombres y mujeres difieren en X, los asiáticoamericanos tienen más de X que los hispanos. Todas estas afirmaciones se refieren a las diferencias en el promedio o la media del desempeño en el rasgo X. Un lector no muy sofisticado puede inferir que todos los del grupo A son superiores a todos los del grupo B, que todos los hombres son diferentes de todas las mujeres, y así sucesivamente. Sin embargo, en general, lo que encontramos es que a) hay una gran variabilidad dentro de cada grupo y b) la distribución de los dos grupos se superpone de manera considerable. La figura 7-12 muestra la **superposición de distribuciones**. Podemos notar que muchos integrantes del grupo B superan la media (promedio) del desempeño de las personas del grupo A; del mismo modo, muchos integrantes del grupo A están por debajo de la media del desempeño de las personas del grupo B.

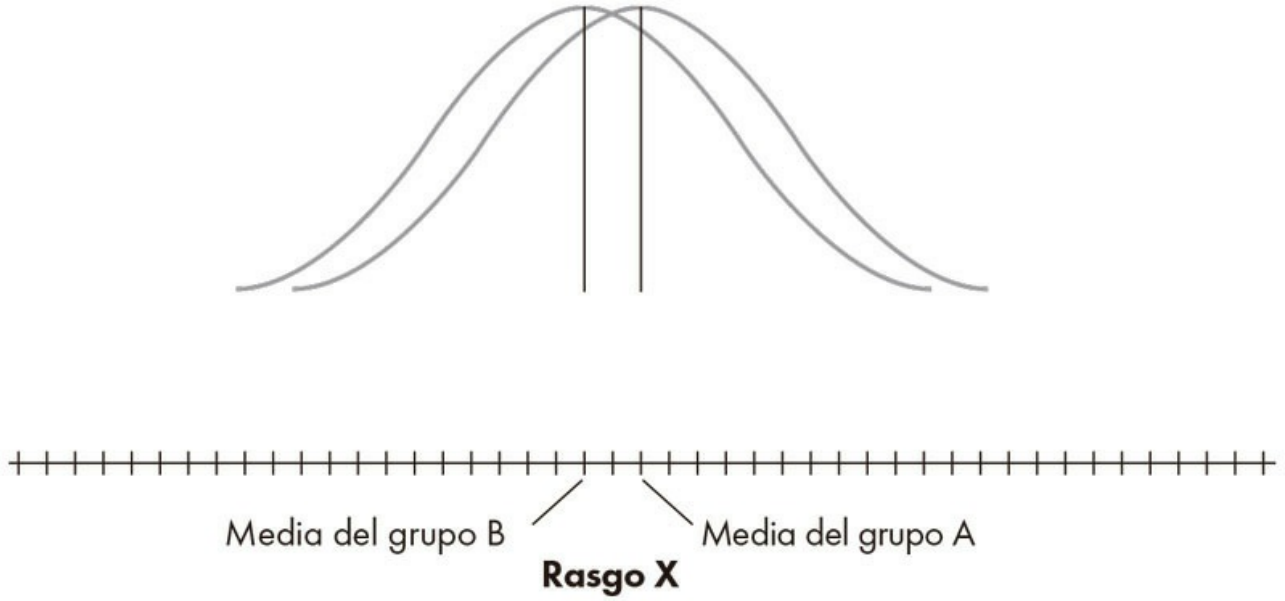


Figura 7-12. Ejemplo de la superposición de distribuciones en el rasgo X en los grupos A y B.

Resumen de puntos clave 7-3

Tres perspectivas importantes en el estudio de las diferencias grupales

1. La superposición de las distribuciones es la regla
2. La diferencia grupal por sí misma no revela su causa
3. Las diferencias pueden cambiar con el tiempo

La situación representada en la figura 7-12 prevalece en prácticamente todos los estudios de diferencias grupales. El método preferido para informar estas diferencias es la desviación estándar (DE) o unidades sigma ([σ]). En el lenguaje de la estadística lo llamamos **tamaño del efecto** (véase Grissom & Kim, 2012, para más detalles). Por ejemplo, el grupo A está .2[σ] arriba del grupo B, por lo que el tamaño del efecto es .2. Este tipo de informe, suponiendo que las [σ] son más o menos equivalentes en los dos grupos, nos permite pensar en la diferencia de una forma similar a la figura 7-12. Desafortunadamente, muchos informes sólo dicen que hay una diferencia (p. ej., el grupo A es superior al grupo B), lo cual es comprensible, quizá, en los medios de comunicación populares, pero también encontramos esto en la literatura psicológica. Señalar que una diferencia es "significativa" o incluso "muy significativa" no evita el fenómeno de la superposición de las distribuciones. Una diferencia entre los grupos puede ser muy significativa y, sin embargo, la distribución aún se superpone de manera considerable. Esto es particularmente cierto cuando el número de casos es muy grande, lo cual es a menudo lo que ocurre en los estudios de diferencias grupales en la inteligencia.

¡Inténtalo!

Con base en tu conocimiento personal acerca de las diferencias de estatura entre hombres y mujeres, dibuja el grado de superposición de las distribuciones de este rasgo.

La segunda perspectiva importante es que la diferencia grupal por sí misma no revela su causa. Tenemos una tendencia natural a inferir que la diferencia entre el grupo A y el grupo B en el rasgo X tiene algo que ver con características inherentes de cada grupo. Éste es un caso clásico de inferencia de la causa a partir de datos correlacionales. Las diferencias grupales, en realidad, son datos correlacionales, pues representan la correlación entre el rasgo X y la pertenencia grupal, que puede ser codificada como 0 y 1 para A y B[4], respectivamente. Sabiendo que los grupos A y B difieren en el rasgo X, no podemos inferir que la diferencia se deba directamente a la pertenencia grupal. De hecho, puede deberse a alguna otra variable que tiene una asociación (correlación) indirecta con la pertenencia grupal. Consideremos un ejemplo que, reconocemos, es tonto. Digamos que hombres y mujeres difieren en el rasgo de personalidad X; nuestra primera inclinación es pensar que existe algo específico en ser hombre o mujer que ocasiona esta diferencia. Sin embargo, el factor causal clave puede ser la estatura. Los hombres tienden a ser más altos que las mujeres. Puede ser el factor "estatura", más que el género, lo que causa la diferencia en el rasgo de personalidad X. ¿Cómo podríamos esclarecer tales relaciones causales? Por lo común, mediante una enorme cantidad de investigaciones. El punto es que el simple informe de una diferencia grupal no nos dice nada acerca de los vínculos causales.

La tercera perspectiva importante es que las diferencias pueden cambiar con el tiempo; es decir, en la actualidad, una diferencia bien establecida hace 30 años puede haberse esfumado o, al menos, disminuido en gran medida. Una diferencia bien establecida hoy puede no durar para siempre. En el caso de varias de las diferencias que describimos, hay evidencia abundante de que han cambiado con el tiempo; en Lee (2002) puede encontrarse un buen ejemplo de cómo algunas diferencias grupales han cambiado en los últimos 30 años.

Con estas tres perspectivas en mente, revisamos las siguientes diferencias grupales en la inteligencia. La mayor parte de la investigación se ha concentrado en las diferencias de acuerdo con el género, la edad, el estatus socioeconómico y la pertenencia racial/ étnica, las cuales retomaremos una por una. Algunas diferencias que antes eran de interés, como las que tenían que ver con la región geográfica o la residencia en zonas urbanas o rurales, ya no lo son, por lo que no las revisaremos.

Diferencias por sexo

¿Hombre y mujeres, o niños y niñas, difieren en inteligencia? Sí y no. En términos de las

puntuaciones totales de pruebas de funcionamiento mental general, las diferencias parecen ser mínimas. En el caso de capacidades más específicas, hay algunas diferencias que vale la pena señalar. La diferencia más pronunciada es la superioridad masculina en algunas pruebas de capacidad espacial. El tamaño del efecto de esta diferencia está entre 0.5 y 0.7. Sobre todo durante los años de desarrollo, las mujeres dejan atrás a los hombres en las habilidades verbales, pero esta diferencia desaparece en la adolescencia tardía. Incluso en el área verbal, hay algunas diferencias más sutiles en el modo en que funcionan hombres y mujeres. Las diferencias por sexo en matemáticas muestran patrones inusualmente complejos según la edad y las subáreas de matemáticas. Por subárea, por ejemplo, las mujeres tienden a ser mejores en cálculo y los hombres, en solución de problemas, pero incluso en este dominio se manifiestan patrones algo diferentes para ciertos tipos de problemas.

Uno de los hallazgos más intrigantes acerca de las diferencias en inteligencia entre los sexos es que la variabilidad es mayor entre los hombres que entre las mujeres. La figura 7-13 ilustra este tipo de diferencia. El efecto práctico de esta diferencia es más evidente en los extremos de la distribución. Hay más hombres que mujeres en los niveles más altos y en los más bajos de la distribución de la inteligencia. Cuando este resultado se combina con cualquier diferencia en el promedio de una capacidad específica, los resultados en los extremos de la distribución pueden ser bastante dramáticos.

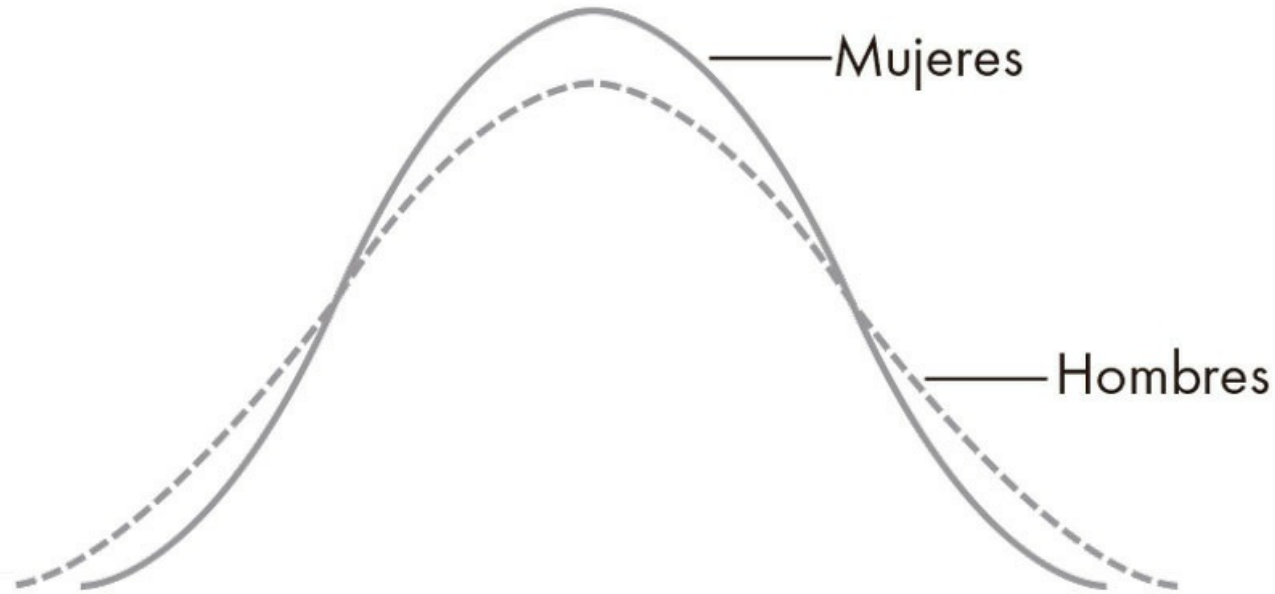


Figura 7-13. Ilustración de la diferencia en variabilidad con el mismo promedio de puntuaciones en hombre y mujeres.

En Halpern, Beninger y Straight (2011) y Hunt (2011, en especial el capítulo 11) se pueden encontrar resúmenes de la vasta literatura sobre las diferencias por sexo en la inteligencia.

Diferencias por edad

Estudiamos las diferencias en la inteligencia por edad graficando las puntuaciones promedio de las pruebas de inteligencia de grupos sucesivos de edad. La figura 7-14 bosqueja las tendencias importantes.

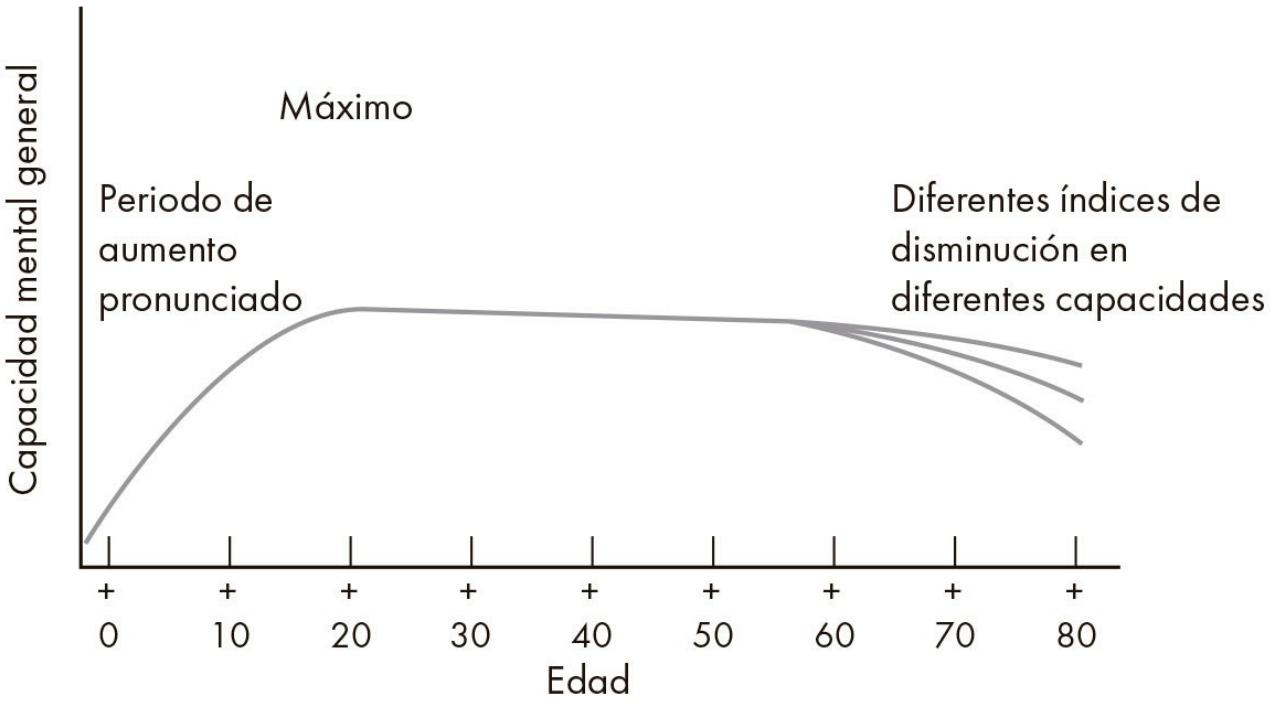


Figura 7-14. Tendencias generalizadas de los cambios por edad en la inteligencia.

El examen de estas gráficas revela las siguientes tendencias. En el caso de la inteligencia general, la curva de crecimiento está muy empinada a lo largo de los primeros 12 años aproximadamente. Después se modera, pero sigue aumentando hasta alrededor de los 20 años. Algunas estimaciones ubican el punto máximo a los 16 años y otras a los 25. Luego, los promedios siguen siendo más o menos los mismos, pero con una ligera disminución alrededor de los 60 años. En este punto, la disminución se hace más notable; en los últimos años, el índice de disminución aumenta.

El punto más importante acerca de las disminuciones en la adultez tardía son los índices diferenciales de disminución de capacidades o pruebas específicas. Las pruebas como vocabulario e información muestran una disminución menor, mientras que las pruebas de memoria a corto plazo y las capacidades perceptual y espacial disminuyen con mayor rapidez. Cuando estudiamos las diferencias por edad en la inteligencia, debemos ser cuidadosos al distinguir entre estudios transversales y longitudinales. En los primeros, los grupos de mayor edad difieren de los más jóvenes no sólo en edad, sino

también en otros factores, como nivel educativo, historias nutricionales. De ahí que los estudios transversales (que son más fáciles de llevar a cabo), por lo general, muestran una disminución mayor en edades avanzadas en comparación con los estudios longitudinales.

Todos los últimos resultados se relacionan con los cambios en los promedios grupales. También está la cuestión de la estabilidad de la inteligencia relativa a lo largo de la vida, la cual, podemos notar, remite al nivel relativo de la inteligencia. En promedio, todos aumentan su inteligencia hasta la adultez temprana y, después, por lo general, disminuye. Sin embargo, ¿qué hay con la ubicación de un individuo dentro de su grupo de edad? ¿El CI alto de un niño de dos años está destinado a ser un CI alto en la edad adulta? ¿El CI de 90 en cuarto año predice un nivel similar de CI en la preparatoria?

Los niveles relativos de inteligencia empiezan a estabilizarse alrededor de los 6 años; esta estabilidad aumenta hasta alrededor de los 16 años, cuando observamos un alto grado de estabilidad, la cual, desde luego, nunca es absoluta. Algunas personas siguen mejorando su posición relativa, mientras que otros decaen. A mayor edad, mayor estabilidad; por ejemplo, podemos hacer una mejor predicción de la inteligencia a la edad de 25 años a partir de una medida tomada a los 16 que de una tomada a los 12; del mismo modo, se puede hacer una mejor predicción a partir de una medida tomada a los 12 que a los 10 años, y así sucesivamente. Además, mientras más cercanas sean las edades en que se toman las mediciones, mayor es la semejanza en las puntuaciones; por ejemplo, una predicción a dos años será mejor que una a cuatro años. Antes de los 6 años, los niveles relativos de inteligencia no son muy estables; así, el CI obtenido en una prueba no es un buen predictor de la inteligencia posterior. En Hertzog (2011), Hoyer y Touron (2003), Hunt (2011) y O'Connor y Kaplan (2003) se pueden encontrar resúmenes de la investigación sobre los cambios por edad en la inteligencia.

Cambios en la población a lo largo del tiempo

Otro modo de pensar acerca de la estabilidad de la inteligencia es en términos de poblaciones enteras. ¿El CI promedio actual es superior al de hace, digamos, 50 años? En apariencia, la respuesta es "sí", y en un grado sorprendente. Los datos pertinentes provienen de los programas de evaluación nacional, en especial los de evaluación militar aplicada a todos sus elementos y de la reestandarización de las pruebas. Varios investigadores señalaron las tendencias, casi como información incidental, a lo largo de los años. James Flynn hizo un trabajo extraordinario al resumir los datos de muchas fuentes de 20 países en los últimos 60 años (Flynn, 1984, 1987, 1994, 1999, 2011). De ahí que a los niveles de CI, que aumentan de manera constante, se les ha denominado "**efecto Flynn**". Algunas fuentes se refieren a los cambios con el nombre de "tendencias profanas" en la inteligencia. Los resúmenes de Flynn muestran cantidades diferentes de aumento en las pruebas que, se supone, miden inteligencia fluida frente a la cristalizada. Las pruebas más relacionadas con la inteligencia fluida (p. ej., razonamiento espacial y con imágenes) muestran un aumento promedio de cerca de 15 puntos por generación.

(No hay una cuantificación exacta del término "una generación", pero suele referirse a un periodo de 20 o 25 años.) Por su parte, las medidas más relacionadas con la inteligencia cristalizada (p. ej., vocabulario y comprensión verbal) muestran un promedio de aumentos de cerca de 9 puntos por generación. Si promediamos estas dos áreas, el aumento es de cerca de 12 puntos por generación, lo que significa que un CI de 100 en nuestros días habría sido un CI de 124 en el tiempo de nuestros abuelos (es decir, hace dos generaciones). En la otra dirección, un CI de 100 en la época de nuestros abuelos hoy sería un CI de 76. Éstas son diferencias muy grandes. Se han propuesto numerosas hipótesis para explicar estos resultados (véase Neisser, 1998), pero ninguna ha encontrado aceptación universal. ¿Se debe más a la educación escolar? ¿A una mejor nutrición? Flynn sugiere que, quizá, las puntuaciones de las pruebas (los CI) aumentan sin ningún cambio real en los niveles de inteligencia o, al menos, que los cambios en la inteligencia subyacente no son tan grandes como los cambios en el CI. Sin duda, la búsqueda de explicaciones del efecto Flynn continuará; desde luego, todos estamos ansiosos por saber si la tendencia seguirá; si esto sigue así en cuatro o cinco generaciones, ¡la persona promedio al final del siglo XXI tendrá un CI de cerca de 160 de acuerdo con las normas actuales! Por otro lado, algunos investigadores sugieren que el aumento se ha detenido y, tal vez, incluso se está revirtiendo.

Diferencias por nivel socioeconómico

Las dos variables previas –edad y sexo– se pueden definir con facilidad, pero el estatus socioeconómico (ESE) es una variable más compleja. Varios estudios lo definen en términos del ingreso familiar, ocupación o nivel educativo, y también se usan combinaciones de estas variables. Además, algunos estudios tratan el estatus socioeconómico como una variable continua, mientras que otros crean grupos de, por ejemplo, ESE alto, medio y bajo.

En cualquier definición que se use, hay una clara relación entre los niveles de inteligencia y el ESE. Cuando éste se representa como una variable continua, la correlación con las puntuaciones de pruebas de inteligencia es de cerca de .30. Muchos investigadores usan cinco grupos de ESE, en cuyo caso las diferencias promedio entre grupos sucesivos son, por lo general, de 5 a 10 puntos de CI. La figura 7-15 presenta una versión generalizada de este resumen. La línea de regresión a través de las medianas de los cinco grupos corresponde a la correlación entre ESE y CI de .30 aproximadamente. La figura también ilustra la superposición de las distribuciones. Podemos notar que hay una tendencia distinta, pero también una considerable superposición.

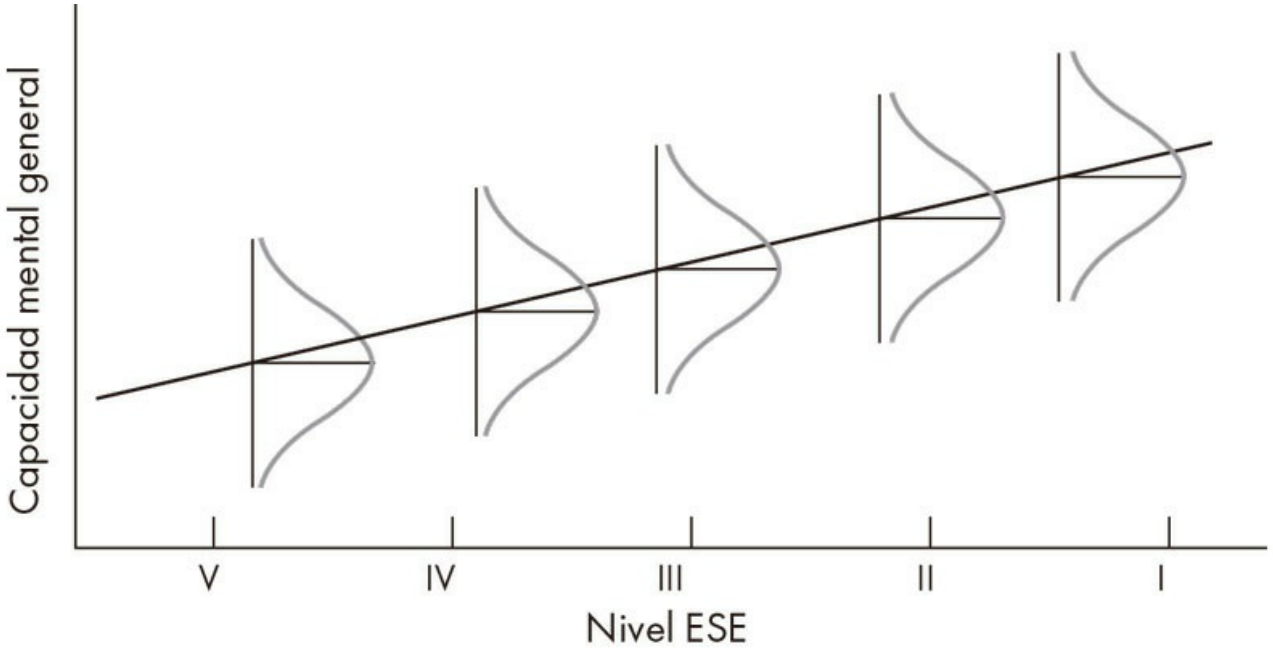


Figura 7-15. Relación generalizada entre CI y el nivel ESE, el cual está representado por grupos.

La razón de la relación entre CI y ESE es rebatida con vehemencia. ¿Las personas más brillantes se mueven a la parte alta de la escala del ESE? ¿O las pruebas son más una medida de la cultura del ESE que de la inteligencia? Mientras este debate acerca de las raíces de las diferencias sigue vivo, pocos se preguntan sobre la magnitud de las diferencias mismas. En Gottfredson (2004), Herrnstein y Murray (1994), Hunt (2011, en especial el capítulo 9), Jensen (1998) y Neisser et al. (1996) se pueden consultar resúmenes sobre las diferencias del ESE. Algunos de estos autores toman rumbos diferentes en la explicación de las diferencias por ESE, pero todas brindan información básica acerca de la magnitud de las diferencias. En Sackett et al. (2009) se puede encontrar un análisis útil de la relación entre ESE y desempeño en las pruebas de admisión a la universidad.

Diferencias por grupo racial/étnico

Aún más discutidas que las diferencias en la inteligencia por estatus socioeconómico son las diferencias por grupo racial/étnico. Sin embargo, igual que con las diferencias por ESE, las controversias se relacionan primordialmente con las causas, pues respecto de la magnitud y dirección de las diferencias, hay un buen consenso. Debemos notar que, debajo de esta discusión, se encuentra la pregunta acerca de la legitimidad del concepto de raza. Algunos autores creen que este concepto no tiene suficiente validez biológica para ser útil, mientras que otros señalan que, incluso si el concepto tuviera cierta validez, los grupos a los que aplicamos los términos de grupo racial o étnico son demasiado

heterogéneos para que tenga sentido usarlos. Por ejemplo, el término hispano abarca personas provenientes de diversas culturas y áreas geográficas. Los asiáticos incluyen a los chinos, japoneses, vietnamitas, camboyanos y muchos otros grupos. Todos estos son buenos puntos; sin embargo, el hecho es que hay una enorme cantidad de investigaciones que informan las diferencias entre blancos, negros, hispanos y asiáticos, así como una inmensa cantidad de otros grupos definidos por su aspecto y/o su origen geográfico. Trataremos de resumir los puntos en los que parece haber acuerdo.

En la mayoría de los informes de investigación, los blancos, específicamente los estadounidenses o caucásicos del oeste de Europa, constituyen el grupo mayoritario con el que se compara a otros grupos. Además, la mayoría de las pruebas que se usan en la investigación tiene a los blancos como el grupo mayoritario en sus normas, por lo que el CI promedio de los blancos es aproximadamente de 100. Por lo general, los negros están en promedio cerca de una desviación estándar (una unidad [σ]) por debajo de los blancos; esto también tiende a ser cierto en capacidades más específicas, aunque la diferencia puede ser un poco menor en las pruebas de ejecución en comparación con las pruebas más verbales. Existe cierta evidencia de que esta diferencia puede hacerse más limitada con el paso del tiempo. Los hispanos tienden a estar más cerca de la media de los blancos en las pruebas de ejecución y no verbales, pero están por debajo en las pruebas verbales por media o una desviación estándar. Desde luego, un tema crucial en la evaluación de muchos hispanos es el idioma en que se aplican las pruebas; los nativos estadounidenses muestran un patrón similar al de los hispanos. Los asiáticos tienden a estar aproximadamente en la media de los blancos en pruebas verbales, mientras que en las no verbales, en especial en las de capacidades espaciales y figurativas, tienden a estar cerca de una desviación estándar por encima de la media de los blancos. La vasta mayoría de las investigaciones sobre los asiáticos implica sólo personas de origen chino o japonés; la investigación con personas de otras culturas o regiones de Asia es muy escasa.

En Herrnstein y Murray (1994), Hunt (2011), Jencks y Phillips (1998), Jensen (1994, 1998), Neisser et al. (1996), Rushton y Jensen (2005) y Suzuki, Short y Lee (2011) se pueden consultar resúmenes de la investigación sobre las diferencias raciales/étnicas en la inteligencia.

Herencia y ambiente

Para quien guste de la controversia, éste es el campo ideal. Pocos temas en cualquier ciencia, no digamos la psicología, provocan un debate tan apasionado como las influencias de la herencia y el ambiente en la inteligencia. Las discusiones han continuado por 100 años al menos, a veces inclinándose en una dirección, a veces, en la otra. No parece tener fin; sin embargo, hay algunas perspectivas bien establecidas y hechos básicos. Resumiremos estos últimos y dejaremos los más inciertos para que otros se hagan cargo de ellos. También debemos notar que el conocimiento en esta área está creciendo a un ritmo vertiginoso con el despliegue del Proyecto del Genoma Humano y sus derivados.

Malentendidos comunes

Empecemos disipando algunos malentendidos comunes. Primero, nadie que haya estudiado con seriedad este tema cree que la herencia o el ambiente por sí mismos determinan por completo la inteligencia. La cuestión ya no es la herencia o el ambiente; todos los académicos concuerdan en que la inteligencia es resultado de una **interacción** de las influencias de la herencia y el ambiente. Estas influencias no son aditivas, sino que son más como una relación multiplicativa. (Véase en el ejercicio 2 al final del capítulo una ilustración del efecto multiplicativo.) Sin importar cuál sea la influencia de la herencia, si el ambiente es por completo negativo, la inteligencia no se desarrollará: el hijo de dos genios criado encerrado en un closet no desarrollará su inteligencia. O aunque enviemos una rana al Colegio Ivy, la rana nunca llegará a ser sabia. Hoy, las preguntas son acerca de las contribuciones relativas de la herencia y el ambiente, y sobre cómo funciona la interacción entre ellos. Cuidado con los escritores que atacan a "los psicólogos que creen que la inteligencia es hereditaria" o a "los psicólogos que creen que la inteligencia está determinada por el ambiente", pues cada una de estas posiciones implica una influencia exclusiva de uno u otro factor. Éstos son hombres de paja, pues ningún psicólogo de hoy cree en alguna de estas posiciones.

Un segundo malentendido común es que los rasgos hereditarios están presentes desde la concepción o el nacimiento, mientras que las influencias ambientales se desarrollan después. Eso no necesariamente es así. Pensemos en la calvicie; tiene una fuerte influencia hereditaria, pero no se manifiesta hasta los 30 o 40 años de edad, o después. El desarrollo del vello facial en los hombres y los senos en las mujeres son, en gran parte, hereditarios, pero no se manifiestan sino en la pubertad. Por otra parte, el síndrome de alcoholismo fetal está determinado por el ambiente (por el ambiente intrauterino) y está presente en el nacimiento. Un tercer malentendido, relacionado con el segundo, es que las influencias hereditarias son permanentes e invariables, mientras que las influencias ambientales son transitorias y variables; incluso los rasgos con una fuerte influencia hereditaria están sujetos a cambios. Por ejemplo, una persona puede tener una

predisposición hereditaria a alguna enfermedad, pero la puede evitar tomando medicamentos. Además, un rasgo puede ser en su mayor parte hereditario, pero la población entera puede ser influida por factores ambientales; por ejemplo, la estatura es en su mayor parte hereditaria, pero la estatura promedio ha aumentado debido a una mejor nutrición y a los servicios de salud. Por otro lado, el daño cerebral debido a un trauma o la amputación de una pierna están determinados por el ambiente; sin embargo, los efectos son permanentes.

Un cuarto malentendido es que la inteligencia puede tener cierto componente hereditario, pero las características de personalidad están determinadas por influencias ambientales, en especial por la familia y otras experiencias tempranas. Sin embargo, la evidencia sugiere que la herencia tiene un papel importante en las características de personalidad, igual que en la inteligencia.

Metodología y términos

La metodología primaria para estudiar las influencias relativas de la herencia y el ambiente sobre la inteligencia es la distancia genética familiar. De especial interés es el estudio de los gemelos idénticos o **monocigóticos** (MC), los cuales, siendo de un mismo óvulo fertilizado, son un laboratorio natural para observar a dos individuos con la misma dotación genética. Dentro de este grupo, el interés se centra en los pares de gemelos criados en ambientes muy diferentes. Es muy difícil conseguir las muestras apropiadas para estos estudios, pues hay relativamente pocos gemelos idénticos, y extremadamente pocos de ellos han sido criados en ambientes separados. De ahí que sea muy pequeño el número de estudios de este tipo dignos de crédito. No obstante, ya ha habido suficientes para tener resultados significativos comprobados. Los estudios de otros grados de parecido familiar también son pertinentes; los hermanos, incluyendo a los gemelos dicigóticos (DC, no idénticos), tienen la mitad de genes en común. También podemos determinar el grado en que se comparte material genético en otras relaciones.

La proporción de varianza de un rasgo que es atribuible a factores genéticos en una población se designa índice de heredabilidad, representado como h^2. La varianza restante ($1 - h^2$) incluye las influencias ambientales y los errores de medición. Entre las influencias ambientales, hay un interés especial en los factores intrafamiliares, lo cual se denomina **varianza familiar compartida** y se representa a menudo como c^2. La pregunta es: ¿qué tan diferentes son los ambientes para los individuos de una misma familia? Cuando abordamos esta cuestión, es de especial importancia recordar que el "ambiente" empieza en la vida intrauterina. También hay varianza ambiental interfamiliar, es decir, las diferencias ambientales entre una familia y otra.

Principales resultados

Aquí identificamos cuatro conclusiones importantes a partir de los estudios sobre la

heredabilidad de la inteligencia. Primero, las estimaciones de la heredabilidad de la inteligencia varían entre .40 y .80. Cuando resumen los resultados de varios estudios, muchos autores usan el punto intermedio de este rango, es decir, .60, como una buena estimación. Algunos autores lo redondean en .50. Segundo, ahora parece bastante sólida la evidencia de que la heredabilidad aumenta con la edad; estimaciones razonables indican que es de .40 a .60 en la juventud, pero las cifras se elevan a .60 o .75 entre los adultos. Sin duda esto va contra el sentido común; pensaríamos que las influencias ambientales mostrarían una importancia creciente con la edad, pero al parecer no es así. Tercero, la mayoría de los estudios definen la inteligencia como una medida de funcionamiento intelectual general, que de manera razonable se interpreta como "g". Los estudios que emplean pruebas de capacidades más específicas sugieren que éstas tienen índices de heredabilidad más bajos. Por ejemplo, en un excelente resumen de la investigación sobre este tema, Plomin y DeFries (1998) concluyeron que la heredabilidad es de cerca de 60% en el caso de la capacidad verbal y de 50% en el de la capacidad espacial. Cuarto, muchos autores comentan acerca de la sorprendentemente pequeña contribución de la varianza interfamiliar; la varianza intrafamiliar parece ser más importante. Además, incluso la varianza intrafamiliar disminuye en influencia conforme aumenta la edad.

Ahora señalemos algunas precauciones al obtener conclusiones relacionadas con la genética de la inteligencia. Primero, casi todos los autores hacen hincapié en que sabemos muy poco sobre los mecanismos por los cuales los genes o el ambiente operan sus respectivas influencias en la inteligencia, aunque ésta sea un área de investigación intensa en este momento (véase Plomin et al., 2003; Plomin et al., 2008). ¿Hay un gen "inteligente" o algunos? Al parecer, no. Anholt y Mackay (2010) señalaron que "a partir de los estudios actuales, parece que las diversas manifestaciones de... 'la inteligencia' son reguladas por muchos genes de efecto pequeño y, tal vez, unos pocos, si los hay, genes de efecto grande" (p. 224). Segundo, las estimaciones de heredabilidad se aplican dentro de poblaciones que comparten una reserva de genes, pues es arriesgado generalizar a otras reservas de genes y es difícil definir con exactitud cuándo nos encontramos con una reserva nueva. Tercero, y relacionado con el segundo punto, podemos notar que la gran mayoría de estudios sobre este tema se han llevado a cabo con poblaciones estadounidenses y de Europa occidental. La referencia estándar sobre la genética de la inteligencia es Behavioral Genetics [Genética de la conducta] (Plomin et al., 2008), ahora en su quinta edición. Anholt y Mackay (2010) también son una fuente útil, aunque se concentran en los animales subhumanos.

Resumen

1. El estudio de la inteligencia cae en cuatro categorías amplias: teorías, medición, diferencias grupales e influencias hereditarias y ambientales.
2. A lo largo de los años, los psicólogos han usado los términos inteligencia, capacidad mental, aptitud y varios más para referirse al rasgo que estudiamos en éste y los siguientes capítulos. Aunque algunas fuentes dicen que los psicólogos no concuerdan en el significado del término inteligencia, en realidad hay un gran acuerdo. La inteligencia se correlaciona con variables prácticas importantes, pero los expertos difieren en la interpretación de la magnitud de las correlaciones.
3. Es importante conocer las teorías de la inteligencia, porque las pruebas actuales se apoyan en gran medida en dichas teorías.
4. La teoría de "g" de Spearman fue la primera teoría formal de la inteligencia. En ella se postula un factor general dominante y una gran cantidad de factores más específicos.
5. Por muchos años, la principal competidora de la teoría de Spearman fue la teoría de las capacidades mentales primarias o teoría multifactorial de Thurstone. En ella se postularon de 5 a 10 factores relativamente independientes. Guilford pensaba que había hasta 180 capacidades relativamente independientes.
6. Los modelos jerárquicos combinan las posturas de un solo factor y las multifactoriales señalando que los factores múltiples pueden formar una jerarquía, en la que "g" ocupa la cima. Los modelos de este tipo más conocidos incluyen el de Vernon, Carroll y Cattell. El modelo de Cattell introduce las nociones de inteligencia fluida y cristalizada.
7. Otro modo de pensar acerca de la capacidad mental es en términos de los modelos de desarrollo. La teoría del desarrollo cognitivo de Piaget ha tenido gran influencia.
8. En años recientes, los modelos del procesamiento de información y biológicos han dominado la literatura de investigación. Entre las teorías más conocidas están las de Jensen, Sternberg y Gardner. La teoría PASS, que incorpora el procesamiento simultáneo y secuencial, es otro modelo de esta categoría. La investigación que emplea tareas cognitivas elementales tiene una especial actividad en nuestros días.
9. Las pruebas actuales se apoyan en su mayor parte en los modelos jerárquicos. Los modelos del desarrollo y del procesamiento de información, aunque atractivos desde muchos puntos de vista, aún no han ejercido una gran influencia práctica en las pruebas de capacidad mental. La investigación sobre la memoria de trabajo empieza a tener cierta influencia en el campo práctico de las pruebas.
10. Cuando se consideran las diferencias grupales en la inteligencia, es importante recordar la regla de la superposición de las distribuciones, el hecho de que las causas de las diferencias pueden ser inaprensibles y que las diferencias pueden cambiar con el tiempo.
11. En el caso de la capacidad mental general, las diferencias por sexo son

insignificantes, pero puede haber algunas diferencias en capacidades más específicas. La mayor variabilidad en los hombres es una diferencia intrigante.

12. La capacidad mental general aumenta rápidamente con la edad hasta la pubertad, luego su ritmo de crecimiento es más moderado y alcanza su punto máximo en la adultez temprana. Al envejecer, capacidades específicas disminuyen a ritmos diferentes.

13. El efecto Flynn describe una tendencia ascendente en el desempeño en pruebas de inteligencia en muchos países en las últimas generaciones. Las razones de esta tendencia no están claras.

14. El estatus socioeconómico tiene una relación moderada con las puntuaciones de las pruebas de inteligencia. Las interpretaciones de la dirección causal difieren notablemente.

15. Hay un consenso razonable acerca de la dirección y magnitud de las diferencias entre varios grupos raciales/étnicos en las pruebas de capacidad mental, pero la controversia importante sigue girando en torno a las razones de estas diferencias.

16. Las estimaciones de la heredabilidad de la inteligencia general están alrededor de .60. Cuando se estudian las influencias hereditarias y ambientales sobre la inteligencia (o cualquier otro rasgo), se deben evitar ciertos malentendidos comunes.

Palabras clave

capacidades mentales primarias
distribuciones superpuestas
efecto Flynn
"*g*"
índice de heredabilidad
inteligencia cristalizada
inteligencia fluida
inteligencias múltiples (IM)
interacción
Jensen
memoria de trabajo
modelo biológico
modelo del procesamiento de información
modelo jerárquico
monocigótico
procesamiento secuencial
procesamiento simultáneo
producción convergente
producción divergente
Spearman
tamaño del efecto
tareas cognitivas elementales (TCE)
teoría bifactorial
teoría de los tres estratos
teoría PASS
teoría triárquica
teorías del desarrollo
teoría multifactorial
teorías de etapas
teorías psicométricas
Thurstone
varianza familiar compartida

Ejercicios

1. Regresa al cuadro 7-2. ¿Cuál de estas capacidades consideras que es importante para el éxito en la universidad?
2. En este capítulo, recomendamos pensar la herencia y el ambiente como una relación multiplicativa. Piensa en el factor genérico y en el ambiental como escalas separadas del 1 al 10. En ambas, 10 equivale a alto o favorable. Llena los espacios en el cuadro para ver el resultado de la relación multiplicativa.

Caso	Herencia	Ambiente	Resultado
1	5	6	_____
2	1	10	_____
3	10	1	_____
4	4	8	_____
5	7	7	_____

3. Compara los factores del estrato II del modelo de tres estratos de Carroll con las capacidades mentales primarias de Thurstone. ¿En dónde concuerdan estos dos modelos? ¿En dónde discrepan?
4. Para observar la diversidad de términos empleados para hablar de "inteligencia", entra al sitio ETS Test Collection (http://www.ets.org/test_link/about) y haz una búsqueda por tema utilizando "intelligence" como palabra clave. Enumera 10 términos que encuentres que describan o nombren las entradas resultantes.
5. Recuerda la distinción entre pensamiento convergente y divergente (p. 181). ¿A qué se asemeja el pensamiento divergente en el modelo de la inteligencia de Carroll (p. 184)?
6.Haz una búsqueda en internet de "working memory tasks" ["tareas de memoria de trabajo"]. Entra a uno o dos sitios para tener idea de qué clase de tareas se trata. Éste es un sitio al que puedes entrar: http://cognitivefun.net/test/4.
7. Lee otra vez acerca de las tareas cognitivas elementales (pp. 185-187). ¿A qué se asemeja el desempeño en estas tareas en el modelo de Carroll?
8. Observa la figura 7-12, la cual muestra distribuciones que se superponen. Estima el tamaño del efecto de estas distribuciones y compara tus estimaciones con las de alguien más.

Notas

[1] Se denominó "simposio", pero en realidad fue sólo una serie de artículos breves dispersos en tres números de la revista, con la promesa editorial de hacer un tratamiento concluyente más tarde, pero esto nunca se llevó a cabo. Así, llamarlo un "simposio" quizá sugiere más de lo que fue en realidad. Aquí hacemos referencia sólo a la

introducción editorial.

2 En el caso de los contenidos, operaciones y productos, respectivamente, la primera versión del modelo de Guilford tenía 4 × 5 × 6 = 120 celdas (Guilford, 1956). La segunda versión tenía 5 × 5 × 6 = 150 celdas (Guilford, 1959b). La última versión, la que describimos aquí, tenía 180 celdas. Varias fuentes podrían citar cualquiera de estas versiones sin mencionar la evolución del modelo.

3 No tiene ninguna relación con James McKeen Cattell, considerado el padre de las pruebas mentales (véase p. 14a»).

4 De hecho, de acuerdo con ciertas suposiciones acerca de las desviaciones estándar, hay una fórmula sencilla que nos permite convertir una diferencia grupal en un coeficiente de correlación (r) y viceversa.

CAPÍTULO 8
Pruebas individuales de inteligencia

Objetivos

1. Enumerar los usos típicos de las pruebas individuales de inteligencia.
2. Describir las características en común de las pruebas individuales de inteligencia.
3. Identificar las principales características del WAIS, incluyendo su estructura, tipos de puntuaciones, estandarización, confiabilidad, validez e interpretación del perfil.
4. Identificar las principales características del Stanford-Binet.
5. Identificar las principales características del PPVT.
6. Describir las principales características y usos del WMS.
7. Describir el concepto de conducta adaptativa y cómo se relaciona con la definición de discapacidad intelectual.
8. Describir las tendencias importantes en la elaboración y uso de las pruebas individuales de inteligencia.

Algunos casos

Bill cursa el sexto grado en la escuela Highland; ha batallado con sus tareas escolares desde segundo grado. Su hermano y hermana mayores son estudiantes sobresalientes y ahora están en universidades prestigiosas. La mamá de Bill se pregunta si Bill no se está esforzando o, quizá, lo hace, pero simplemente no tiene el mismo grado de agudeza mental que sus hermanos mayores. Tal vez existe un problema de aprendizaje que requiere un método pedagógico diferente. Los padres de Bill le piden al psicólogo escolar que evalúe su nivel de capacidad mental.

De acuerdo con sus amigos, la señora Kelly, de 75 años de edad, solía tener una "gran agudeza", pero parece ya no tenerla. En parte, esto se debe a fallas en su memoria, pero también hay otros signos que la delatan. La señora Kelly es enviada con un psicólogo clínico especializado en casos geriátricos para que se realice una evaluación general de su funcionamiento mental. Quizá los cambios sean bastante normales para su edad, o quizá no.

En un accidente automovilístico reciente, Sue sufrió una conmoción cerebral grave. ¿Hay evidencia de que el accidente haya afectado su nivel de inteligencia? ¿Cómo podemos responder esta pregunta?

Todos estos casos ilustran situaciones en que el psicólogo podría usar una prueba de capacidad mental de aplicación individual. En este capítulo, exploraremos los tipos de pruebas que los psicólogos usan a menudo en casos como éstos.

Usos y características de las pruebas individuales de inteligencia

Las pruebas de inteligencia de aplicación individual son artículos de primera necesidad en los campos de la psicología clínica, escolar y de orientación. En muchos casos, el psicólogo necesita alguna medida de la capacidad mental general del cliente. Una prueba de inteligencia, por lo común, se usa junto con otras fuentes de información, como las entrevistas u otros tipos de pruebas, dependiendo de la naturaleza del problema. Las otras pruebas pueden ser de personalidad o de otras funciones mentales más específicas. Sin embargo, una medida de capacidad mental general a menudo servirá como una fuente esencial de información. Las pruebas individuales de inteligencia también desempeñan una función importante en la investigación; en algunos casos, directamente sobre la naturaleza de la inteligencia. No obstante, debido a su importancia, también se requiere una medida de capacidad mental general en otras áreas de la investigación, por ejemplo, sobre los ambientes familiares o las características personales.

Figura 8-1. Disposición típica del espacio para la aplicación de una prueba individual de inteligencia.

Las pruebas de inteligencia de aplicación individual tienen varias características en común, que será útil identificar antes de describir las pruebas específicas. Enumeramos ocho de estas características; las más obvia es que son de **aplicación individual.** Hay un examinador y un examinado. En la nostálgica terminología de los primeros días de la psicología experimental, los manuales de las pruebas a veces se referían al examinado como el "sujeto", palabra que ahora no es aceptada. El examinador presenta preguntas, reactivos o estímulos al examinado, y éste responde de alguna manera: puede ser oral (p. ej., definiendo una palabra), manual (p. ej., armando un rompecabezas) o señalando (p. ej., con el dedo o incluso con los ojos).

Segundo, la aplicación de estas pruebas requiere de un entrenamiento avanzado. Al observar la aplicación de una de estas pruebas, se tiene la impresión de que es fácil, casi como una conversación casual. Sin embargo, esta apariencia se logra con mucha práctica y un estudio cuidadoso del manual de la prueba. Es como un ballet o un partido de futbol bien ejecutado; se ve fácil, pero requiere horas de práctica, instrucción y esfuerzo.

Tercero, estas pruebas suelen abarcar un muy amplio rango de edades y capacidades; por ejemplo, una sola prueba puede abarcar edades de 3 a 16 años, y dentro de ese rango, del nivel de retraso mental hasta la genialidad. Por lo tanto, dentro de este rango, los reactivos suelen variar desde los más fáciles hasta los más difíciles; sin embargo, un solo examinado realiza únicamente algunos de estos reactivos. El manual de la prueba indica las **reglas de inicio** y **de discontinuación** para determinar qué reactivos presenta efectivamente el examinador. Por ejemplo, la regla de inicio puede decir: "Empiece con los reactivos típicos para personas un año más jóvenes que el examinado", mientras que la regla de interrupción puede decir: "Discontinúe la prueba cuando el examinado falle en cinco reactivos consecutivos". El examinador debe conocer en detalle estas reglas para lograr una aplicación estandarizada.

Cuarto, el examinador debe *establecer rapport* con el examinado. ***Rapport*** es un término semitécnico del campo de las pruebas que se refiere a una relación cálida y confortable entre examinador y examinado. Para establecer *rapport*, el examinador puede necesitar dedicar cierto tiempo sólo a platicar con el individuo antes de que empiece la prueba. Es importante mantener esta relación durante toda la aplicación.

Quinto, la mayoría de las pruebas de inteligencia aplicadas de manera individual usa un formato de respuesta libre en vez de uno de respuesta cerrada. Como veremos en nuestra revisión de pruebas específicas, algunos reactivos pueden emplear este último formato, y en algunas pruebas breves, pueden ser de opción múltiple en su totalidad. La regla general para las pruebas de inteligencia que más se usan es emplear reactivos de respuesta libre; sin embargo, se puede percibir un movimiento en favor del uso de reactivos de respuesta cerrada.

Sexto, las pruebas individuales requieren, por lo común, la **calificación inmediata de los reactivos.** Es decir, el examinador califica cada respuesta conforme el examinado la emite, por lo que es necesario el uso apropiado de las reglas de inicio y discontinuación. Por ejemplo, si la aplicación se discontinúa después de cinco fallas consecutivas, el examinador debe saber de inmediato cuándo ocurre esta quinta falla. Además, ya que la

mayoría de pruebas usa un formato de respuesta libre, el examinador tiene que decidir de inmediato si pedir que el examinado aclare una respuesta en particular. La capacidad del examinador para realizar la calificación inmediata es una de las razones más importantes del entrenamiento avanzado en el uso de estas pruebas.

Séptimo, la aplicación de las pruebas individuales de inteligencia más usadas, por lo general, requiere cerca de una hora. No hay ningún tiempo exacto especificado, porque cada examinado responde diferentes reactivos y la selección exacta de éstos se determina durante la aplicación real. Algunos examinados pueden terminar la prueba incluso en 45 min, mientras que otros pueden necesitar hasta 90 min. Una hora es el promedio, pero no se incluye en ella el tiempo para resumir las puntuaciones o convertir las puntuaciones naturales en estándar. Hay algunas pruebas individuales que se aplican en menos tiempo, pues están diseñadas para contestarse en 15 min. Describimos algunas de éstas más adelante en este capítulo.

Por último, las pruebas individuales de inteligencia brindan la oportunidad de observar la ejecución del evaluado, independientemente de la calificación formal de la prueba. Ésta es una de las grandes ventajas de las pruebas de aplicación individual en comparación con las de aplicación grupal. Por ejemplo, el examinador puede observar la manera en que el examinado se acerca a la tarea, notar gestos o inferir algo sobre su personalidad. El registro de la prueba, donde el examinador anota las respuestas, por lo general ofrece espacio para tomar notas sobre tales observaciones. El cuadro 8-1 muestra ejemplos de observaciones que podrían hacerse durante la aplicación de una prueba. Estas observaciones no forman parte de la calificación formal de la prueba, pero pueden ser útiles al preparar el informe sobre el individuo, el cual suele ir más allá de las simples puntuaciones.

Cuadro 8-1. Ejemplos de los comentarios que un examinador registra durante la aplicación de una prueba individual de inteligencia

JK* parecía muy prudente y metódico al acercarse a las tareas. A menudo, articulaba un "plan" para realizar un reactivo.
EB buscó muchos pretextos para su desempeño... incluso cuando sus respuestas eran correctas. Parecía tener una actitud defensiva.
LN a menudo pedía que le repitiera una pregunta. Puede ser un problema de atención o de audición. Necesita hacerse una revisión de su capacidad auditiva.
BV tuvo problemas excepcionales para concentrarse. Sus ojos recorrían todo el salón. Se distraía con facilidad con el más mínimo ruido del vestíbulo.

* Iniciales del examinado.

Resumen de puntos clave 8-1

Ocho características en común de las pruebas de inteligencia de aplicación individual

1. Aplicación individual

2. Requiere un entrenamiento avanzado para aplicarla
3. Amplio rango de edades y capacidades (con reglas de inicio y de discontinuación)
4. Establecimiento del rapport
5. Formato de respuesta libre
6. Calificación inmediata de los reactivos
7. Cerca de una hora de aplicación
8. Oportunidad de observación

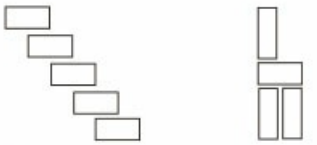
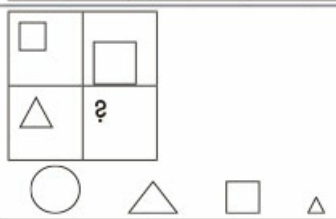
Reactivos típicos en una prueba individual de inteligencia

¿Cómo es que las pruebas de inteligencia de aplicación individual intentan medir la inteligencia? Desde un punto de vista práctico, esta pregunta se traduce como: ¿Qué clase de reactivos tienen estas pruebas? Más adelante en este capítulo, examinaremos pruebas específicas, pero primero consideremos algunos de los reactivos típicos que podemos encontrar en varias pruebas.

Empezaremos tratando de identificar las tareas que, creemos, indican una conducta inteligente. ¿Qué clase de cosas pueden hacer las personas que consideramos inteligentes o brillantes y que no pueden hacer las personas que consideramos intelectualmente torpes o lentas? Necesitamos encontrar tareas que no dependan demasiado de las experiencias culturales. Hay dos enfoques comunes para este propósito. Uno consiste en usar material muy novedoso, que prácticamente nadie haya experimentado antes; otro método es usar material muy común, que casi cualquiera haya experimentado, al menos dentro de una cultura definida de manera amplia, como la "América cotidiana".

El cuadro 8-2 enumera los reactivos que, por lo común, aparecen en las pruebas individuales de inteligencia. Examinemos cada tipo de reactivo y tratemos de determinar qué funciones mentales demanda cada uno. Después de examinar los reactivos, podemos notar las siguientes advertencias respecto de estos ejemplos.

Cuadro 8-2. Ejemplos de reactivos incluidos en pruebas individuales de inteligencia

Categoría	Ejemplos	Comentario
Vocabulario	• ¿Qué significa enojado? • ¿Qué significa arrogante?	Los reactivos de vocabulario son muy comunes en las pruebas de inteligencia. El vocabulario suele tener correlaciones altas con las puntuaciones totales basadas en reactivos de muchas clases. Algunas pruebas consisten totalmente en reactivos de vocabulario.
Relaciones verbales	• ¿Qué es lo opuesto de tardío? • ¿En qué se parece un autobús a un carro? • El padre es a su hijo lo que la madre es a...	Esta categoría incluye antónimos, semejanzas, analogías y otros reactivos que se ocupan de las relaciones entre palabras y conceptos.
Información	• Muéstrame tu codo. • ¿Cuántos días hay en una semana? • Di el nombre de un planeta de nuestro sistema solar que no sea la Tierra.	Es importante conseguir reactivos que no estén demasiado ligados a lo cultural o sean demasiado dependientes de la experiencia escolar. El énfasis está en la información común, cotidiana.
Comprensión del significado	• Haz una oración usando estas palabras: el, Lyz, carro, condujo. • ¿Por qué hay límites de velocidad en las carreteras? • El examinado lee un párrafo y luego se le pide un resumen del tema principal.	Estos reactivos se ocupan de significados diferentes de los de palabras aisladas. Los reactivos hacen hincapié en relaciones, conceptos y vínculos, por lo general, de naturaleza verbal.
Aritmética	• Jim compró dos lápices a 10 pesos cada uno. ¿Cuánto pago? • Jim compró cinco lápices a 12 pesos cada uno y dos libretas a 80 pesos cada una. ¿Cuánto pagó?	Éstos son problemas buenos, que utilizan palabras pasadas de moda. Los reactivos evitan cálculos muy complicados (p. ej., $13/4 \times 2/3$). Los reactivos se concentran en la manipulación de números bastante sencillos.
Memoria a corto plazo	• Escuche y después repita los números que digo: 9-4-7-2-6. • Escuche: perro, casa, vaca, mesa. (Pausa.) ¿Cuál fue la segunda palabra que dije?	El primer reactivo se denomina **retención de dígitos**. Puede emplear cualquier cantidad de dígitos y también puede usarse con dígitos que se repiten en orden inverso. Es evidente que las listas pueden volverse demasiado largas. Algunos reactivos requieren la repetición inmediata, mientras que otros pueden demorar la respuesta por varios min.
Patrones de formas	Usa las piezas para construir una figura como ésta. 	Hay una gran variedad de reactivos que emplean rompecabezas, tableros con formas y cubos. Muchos de estos reactivos eran el modelo de pruebas enteras elaboradas en los inicios del campo de las pruebas, p. ej., Diseño con cubos de Kohs, Laberintos de Porteus y el Tablero de formas de Seguin.
Psicomotor		Estos reactivos suelen ser de velocidad. Este ejemplo requiere sólo un renglón, pero las pruebas reales pueden tener hasta 20 renglones como éste. Las tareas básicas son sencillas; sólo se requiere coordinación ojo-mano y concentración. Otro ejemplo es comparar columnas de números.
Matrices		Las matrices se han vuelto cada vez más populares como reactivos de razonamiento no verbal. Por lo común se presentan como reactivos de opción múltiple y pueden llegar a ser muy complejas.

Éstos son sólo ejemplos de los tipos de reactivos que encontramos en las pruebas individuales de inteligencia; también podemos encontrar otros tipos, pero éstos son los que aparecen con mayor frecuencia. Desde luego, cada tipo de reactivo puede adaptarse a niveles de dificultad muy diferentes. Estos ejemplos no se han sometido a una prueba empírica ni a una revisión editorial; algunos de ellos pueden resultar ser reactivos pobres. Además, no especificamos las reglas de calificación de estos reactivos; por ejemplo, ¿qué aceptaremos como una definición adecuada de "arrogante"? ¿Daremos puntos extra por la velocidad de la respuesta a los reactivos de aritmética o sólo los calificaremos como correctos o incorrectos? Las reglas de calificación se convierten en parte del reactivo.

¡Inténtalo!

Toma alguna de las categorías de reactivos que aparecen en el cuadro 8-2 y construye un reactivo que pienses que pueda medir la capacidad mental de un niño de 10 años de edad.

Escalas Wechsler: panorama general

Introducción histórica

Como señalamos más adelante en nuestra descripción del Stanford-Binet, esta prueba fue por muchos años el método más popular para evaluar la inteligencia. Aunque originalmente se diseñó para aplicarse a niños, también se usó con adultos, pero David Wechsler, psicólogo clínico que trabajaba en el Hospital Bellevue de la ciudad de Nueva York, no estaba satisfecho con la orientación infantil del Stanford-Binet, ni con el hecho de que produjera sólo una puntuación global. Wechsler trabajaba principalmente con adultos; además, quería puntuaciones separadas para lo que le parecía ser manifestaciones separadas de la capacidad mental. Creó lo que después llegó a conocerse como la *Wechsler Bellevue Intelligence Scale* [Escala Wechsler Bellevue de Inteligencia], cuya primera publicación apareció en 1939. La escala Wechsler-Bellevue tocó una cuerda que tuvo una gran resonancia entre los clínicos. En 1955 se hizo una revisión de la prueba y reapareció con el nuevo nombre de **Wechsler Adult Intelligence Scale** [Escala Wechsler de Inteligencia para Adultos] (**WAIS**), el cual se conserva hasta nuestros días.

En un giro inesperado, después de expresar su insatisfacción con la orientación infantil del Stanford-Binet, Wechsler creó una extensión del WAIS hacia edades inferiores: la **Wechsler Intelligence Scale for Children** [Escala Wechsler de Inteligencia para Niños] (**WISC**) para edades de 6 a 16 años, que apareció por primera vez en 1949. La estructura del WISC, por lo general, sigue la del WAIS. Más tarde, el WISC fue llevado todavía más abajo con la publicación de la **Wechsler Preschool and Primary Scale of Intelligence** [Escala Wechsler de Inteligencia para Preescolar y Primaria] (**WPPSI**) para edades de 21 a 27 meses, que se publicó por primera vez en 1967. En la figura 8-2 aparece una línea del tiempo con las ediciones sucesivas del WAIS, WISC y WPPSI.

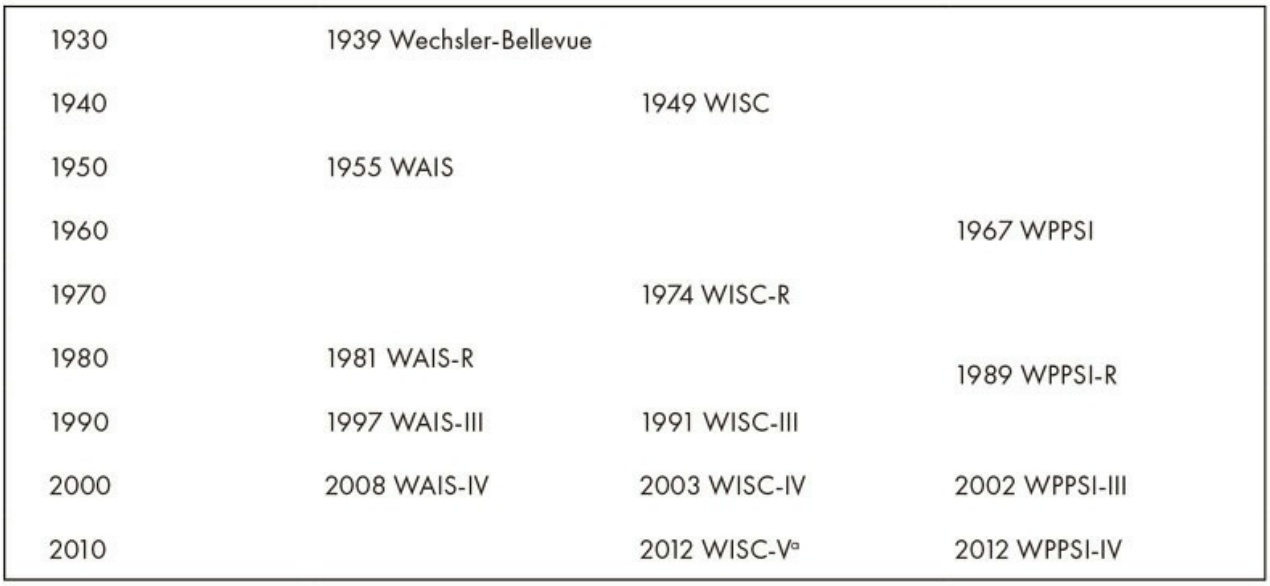


Figura 8-2. Línea del tiempo con las ediciones originales y revisadas del WAIS, WISC y WPPSI.

David Wechsler murió en 1981, pero su nombre se sigue usando en las nuevas ediciones de sus pruebas e, incluso, en pruebas por completo nuevas elaboradas después de su muerte. El cuadro 8-3 muestra las numerosas pruebas Wechsler actuales, algunas de las cuales son pruebas de aprovechamiento más que de inteligencia (véase en el capítulo 11 la información sobre el WIAT).

Cuadro 8-3. La familia Wechsler de pruebas

Título de la prueba	Fecha de publicación	Acrónimo
Escala Wechsler de Inteligencia para Adultos, 4a. ed.	2008	WAIS
Escala Wechsler de Inteligencia para Niños, 4a. ed.	2003	WISC
Escala Wechsler de Inteligencia para Preescolar y Primaria, 4a. ed.	2012	WPPSI
Escala de Memoria Wechsler, 4a. ed.	2009	WMS
Wechsler Abbreviated Scale of Intelligence, 2a. ed.	2011	WASI
Wechsler Nonverbal Scale of Ability	2006	WNV
Wechsler Individual Achievement Test, 3a. ed.	2009	WIAT
Wechsler Test of Adult Reading	2001	WTAR
Wechsler Fundamentals: Academic Skills	2008	(ninguno)

Los psicólogos suelen pronunciar los acrónimos de las principales pruebas Wechsler; usar esta jerga es como una insignia de madurez profesional, aunque una muy menor. Aquí presentamos una breve guía de la pronunciación de las pruebas Wechsler (Wex-ler): WAIS rima con “face”; WISC, con “brisk”; y WPPSI, con “gypsy”.

El concepto de inteligencia de Wechsler

A lo largo de las diversas ediciones de sus escalas de inteligencia, Wechsler definió de manera consistente la inteligencia como "la capacidad total o global del individuo para actuar de manera propositiva, pensar racionalmente y desenvolverse eficazmente en el ambiente." (Wechsler, 1958, p. 7). También hizo hincapié en que la inteligencia implicaba más que la capacidad intelectual, aunque "la capacidad para hacer trabajo intelectual es un signo importante y necesario de la inteligencia general" (Wechsler, 1958, p. 12). La inteligencia general o, de manera más precisa, la conducta inteligente depende de variables como la "persistencia, empuje, nivel de energía, etc." (Wechsler, 1949, p. 5). En Wechsler (1974) se puede encontrar un tratamiento más extenso de sus puntos de vista. Como señalamos antes, Spearman hizo observaciones similares acerca de la naturaleza de la inteligencia; por desgracia, éstas se pierden a menudo en las descripciones de las ideas de Wechsler y Spearman acerca de la inteligencia. Wechsler esperaba que su combinación de pruebas, que en breve examinaremos, captara el "conjunto" de capacidades y rasgos. En el cuadro 8-4 aparecen las definiciones clásicas de inteligencia tanto de Wechsler como de Binet.

Cuadro 8-4. Definiciones de inteligencia de Binet y Wechsler

"la capacidad total o global del individuo para actuar de manera propositiva, pensar racionalmente y desenvolverse eficazmente en el ambiente." (Wechsler, 1958, p. 7)
"Nos parece que en la inteligencia hay una facultad fundamental, cuya alteración o falta es de la mayor importancia para la vida práctica. Esta facultad es el juicio, también llamado buen sentido, sentido práctico o intuitivo, la facultad de adaptarse uno mismo a las circunstancias. Juzgar bien, comprender bien, razonar bien son actividades esenciales de la inteligencia." (Binet, 1905) [a]

[a] Esta afirmación apareció originalmente en un artículo de Binet titulado "New Methods for the Diagnosis of the Intellectual Level of Subnormals", publicado en *L'Année Psychologique* en 1905. Aquí usamos la traducción de E. S. Kite en Binet y Simon (1916).

Escala Wechsler de Inteligencia para Adultos –IV

Entre las pruebas de inteligencia de aplicación individual, la Escala Wechsler de Inteligencia para Adultos (WAIS) es la que más se usa con propósitos aplicados y de investigación. El conocimiento que los psicólogos tienen de ella y su influencia en el campo de las pruebas es, en verdad, sobresaliente. Por lo tanto, vale la pena examinar esta prueba en detalle.

Estructura y aplicación

Una de las características distintivas del WAIS es su estructura. Las primeras pruebas tipo Binet, como se describió en nuestro tratamiento de la historia de las pruebas, producían una sola puntuación general. Desde luego, estas pruebas incluían numerosos tipos de reactivos, por ejemplo, vocabulario, memoria a corto plazo, habilidades psicomotrices, y así sucesivamente, pero estos reactivos diversos sólo contribuían a una medida global de inteligencia. En contraste, Wechsler creía que las diferentes manifestaciones de la inteligencia merecían sus propias puntuaciones al mismo tiempo que se podían resumir en una sola puntuación general. Así surgió la estructura básica de las escalas Wechsler de inteligencia: varias subpruebas, algunas puntuaciones intermedias de resumen y una sola puntuación total. En las primeras tres ediciones del WAIS, las puntuaciones intermedias de resumen fueron las de Inteligencia Verbal y de Ejecución (hablaremos más de esto en un momento). WAIS-IV abandona esta distinción tradicional y usa cuatro índices como puntuaciones de resumen intermedias, que se sitúan entre las puntuaciones de las subpruebas y el **CI Total**. El cuadro 8-5 muestra la organización de las subpruebas en los cuatro índices.

Cuadro 8-5. Organización del WAIS-IV: puntuaciones totales, de índices y de subpruebas

	Puntuación total (CI Total)			
Puntuaciones de índices:	**Comprensión verbal**	**Razonamiento perceptual**	**Memoria de trabajo**	**Velocidad de procesamiento**
Subpruebas				
Semejanzas	X			
Vocabulario	X			
Información	X			
Comprensión	Xs			
Diseño con cubos		X		
Matrices		X		
Rompecabezas visual		X		
Peso figurado		Xs		
Figuras incompletas		Xs		
Retención de dígitos			X	
Aritmética			X	
Sucesión de números y letras			Xs	
Búsqueda de símbolos				X
Claves				X
Cancelación				Xs
(s = prueba suplementaria)				

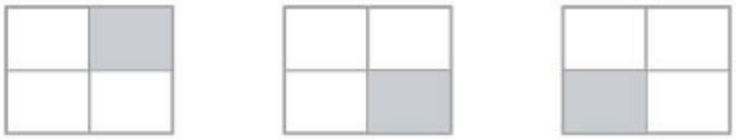
El cuadro 8-6 enumera los nombres de las subpruebas del WAIS-IV junto con una breve descripción de la naturaleza de las tareas. La figura 8-3 ofrece ejemplos de reactivos de algunas subpruebas, pero *no* forman parte de ninguna versión real del WAIS; simplemente ilustran de una manera muy general la naturaleza de dichas tareas. Podemos notar la variedad de tareas que se presentan, lo cual concuerda con la fuerte creencia de Wechsler acerca de que la inteligencia es multifacética y su medición requiere muchos métodos diferentes. Las subpruebas "suplementarias" se aplican en casos donde una de las otras subpruebas se ha "inutilizado", es decir, algo salió mal en su aplicación.

Cuadro 8-6. Descripciones de las subpruebas del WAIS-IV

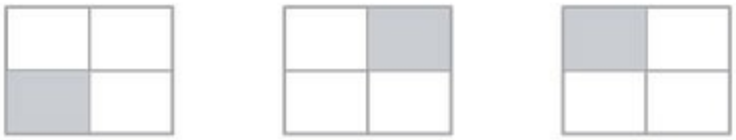
Subprueba	**Descripción**
Diseño con cubos	Con un conjunto de cubos, recrear patrones mostrados en una tarjeta
Semejanzas	Describir la base de la semejanza entre dos cosas
Retención de dígitos	Repetir series de dígitos presentados. Tiene formatos de orden directo y orden inverso
Matrices	Elegir imágenes que completen matrices presentadas de manera visual
Vocabulario	Dar la definición de una palabra
Aritmética	Resolver problemas aritméticos presentados de manera oral

Búsqueda de símbolos	Observar con rapidez un conjunto de símbolos para determinar si ahí aparece cierto símbolo
Rompecabezas visual	Elegir de seis figuras las tres que pueden formar cierta figura completa
Información	Conocimiento de varios hechos simples de historia, geografía, etc.
Claves	Escribir con rapidez símbolos apareados con números con base en una clave que muestra las parejas de símbolos y números
Sucesión de números y letras	Repetir series de números y letras en orden ascendente y alfabético, respectivamente, después de oírlas revueltas
Peso figurado	Determinar las figuras apropiadas que se necesitan de un lado de una báscula para igualar las figuras del otro lado; distintas figuras tienen pesos diferentes
Comprensión	Dar explicaciones acerca de prácticas o eventos comunes
Cancelación	Marcar rápidamente con una diagonal en un conjunto de símbolos aquellos que se muestran en cierto conjunto de símbolos
Figuras incompletas	Identificar la parte faltante de un dibujo

Matrices



Elige el cuadro que completa la serie



1 2 3

4 5

Figuras incompletas

¿Qué falta en este dibujo?

Retención de dígitos

El examinador dice: 5 – 2 – 9
El examinado repite.

El examinador dice: 7 – 1 – 3 – 6 – 9 – 2 – 5 – 4
El examinado repite.

Información

¿Cuál es la distancia en millas de Chicago a Los Ángeles?

Comprensión

¿Por qué tenemos límites de velocidad en las carreteras?

Sucesión de números y letras

El examinador dice: L – 5 – B – 2

El examinado dice: (los números en orden ascendente y las letras en orden alfabético):
2 – 5 – B – L

Figura 8-3. Ejemplos ficticios de algunos tipos de reactivos de las escalas Wechsler.

Los cuatro índices son: Comprensión verbal, Razonamiento perceptual, Memoria de trabajo y Velocidad de procesamiento. Cada uno de ellos deriva de una combinación de 2 o 3 subpruebas, como se indica en el cuadro 8-5. Los índices se obtuvieron del análisis factorial de las subpruebas de Wechsler; numerosos estudios han examinado la estructura factorial de las escalas Wechsler y, por lo común, se identifican tres o cuatro factores. La editorial, al final, adoptó el modelo de cuatro factores, en parte, dependiente de la introducción de ajustes modestos en el conjunto completo de subpruebas.

WAIS-IV mantiene la mayoría de las subpruebas de la edición previa y, de hecho, de la

primera edición del WAIS (Wechsler-Bellevue). Para los usuarios experimentados, es difícil imaginar un Wechsler sin, digamos, Vocabulario, Semejanzas, Diseños con cubos y Retención de dígitos. Sin embargo, cada edición trae algunos cambios. En particular, WAIS-IV introduce las subpruebas Peso figurado, Rompecabezas visual y Cancelación, y deja fuera Ensamble de objetos y Ordenamiento de dibujos; además, hace algunos ajustes respecto de qué subpruebas se consideran suplementarias.

WAIS-IV sigue buscando medir la inteligencia y no tiene ninguna duda respecto de usar justo ese término. Los párrafos de apertura del Manual técnico e interpretativo (Wechsler, 2008a, p. 1) hacen referencia a "inteligencia", "funcionamiento intelectual" y "capacidad intelectual general". Esto presenta una continuidad interesante en un momento en que muchos autores y editoriales de pruebas luchan por presentarse con términos alternativos.

La aplicación del WAIS sigue el patrón bosquejado arriba de las características en común de las pruebas de inteligencia de aplicación individual. El examinador y el examinado se sientan en una mesa uno frente al otro. El examinador tiene los materiales de la prueba y la hoja de registro de las respuestas sobre la mesa. Dedica cierto tiempo a "establecer el *rapport*". En cada subprueba, hay una regla de inicio que determina dónde empezar; si hay una falla inicial, el examinador aplica reactivos más fáciles; para ello, tiene que calificar cada reactivo conforme lo responde el examinado. Algunas calificaciones son muy sencillas, como las de opción múltiple de Matrices o las de Retención de dígitos, pero otras subpruebas (como Vocabulario) requieren del juicio del examinador. Además, en algunas subpruebas, el examinador debe cronometrar las respuestas. La aplicación de las subpruebas prosigue hasta que se cumpla la regla de discontinuación; entonces se continúa con la siguiente subprueba. El tiempo de aplicación suele oscilar entre 45 y 90 min.

Aquí hay un desarrollo interesante. Una de las características distintivas de las pruebas de inteligencia de aplicación individual, como señalamos antes, es el uso de un formato de respuesta libre: el examinador hace una pregunta o presenta un estímulo, y el examinado construye una respuesta. De hecho, la mayoría de los reactivos del WAIS son de este formato. Sin embargo, es interesante observar que varias de las subpruebas más recientes emplean un formato de opción múltiple. Por ejemplo, Matrices, introducida en el WAIS-III y mantenida en el WAIS-IV, y la nueva subprueba Peso figurado son por completo de opción múltiple; la nueva subprueba Rompecabezas visual, que requiere de elegir tres de seis elementos, también es, de hecho, de opción múltiple. ¿Este movimiento sugiere una tendencia que se está desarrollando en las pruebas de inteligencia de aplicación individual?

En los primeros días, las puntuaciones separadas de las escalas Verbal y de Ejecución y un perfil de subpruebas distinguieron las pruebas Wechsler del Stanford-Binet y su puntuación global única. Sin embargo, como describimos más adelante, el Stanford-Binet ahora usa un método de puntuaciones múltiples. Como veremos en el capítulo 9, el abuelito de las pruebas de capacidad mental de aplicación grupal, la prueba Otis, ha seguido una ruta similar. Es claro que alguna versión del modelo jerárquico de inteligencia

se ha impuesto. La figura 8-4 muestra la estructura jerárquica del WAIS-IV; ¿se ve como los modelos jerárquicos que examinamos en el capítulo 7? Podemos apostar a que sí. La lección que hay que aprender es: si queremos comprender la estructura de las pruebas modernas de inteligencia, es mejor conocer cómo funciona el análisis factorial.

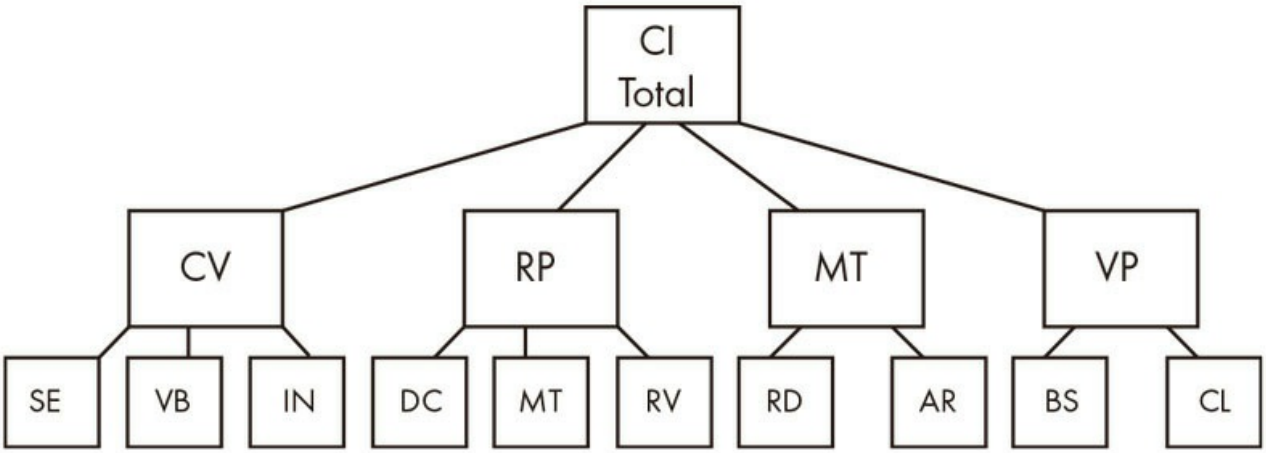


Figura 8-4. Organización gráfica del WAIS-IV (sin incluir las pruebas suplementarias).

Además de las puntuaciones de subpruebas, índices y CI Total, WAIS-IV ofrece *Puntuaciones de proceso*, que se obtienen de algunas subpruebas que se califican de manera especial. Por ejemplo, la puntuación de Retención de dígitos se basa en reactivos en que se pide repetir los dígitos en el mismo orden en que se presentan o en orden inverso. Sin embargo, se pueden generar puntuaciones separadas para las dos partes de la prueba. La manera ordinaria de calificar Diseño con cubos incluye puntos de bonificación que se basan en qué tan rápido termina un diseño el examinado. Sin embargo, también se puede generar una puntuación sin emplear tales puntos de bonificación. Estas "puntuaciones de proceso" se desarrollaron en la literatura clínica con el paso de los años y no son por completo nuevas en el WAIS, pero en el WAIS-IV se les ha dado mayor importancia que en las ediciones anteriores.

Como señalamos antes, WAIS-IV abandona el CI Verbal y el CI de Ejecución (CIV, CIE), que estuvieron consolidados por mucho tiempo como el mecanismo estándar para informar las puntuaciones, en favor de las cuatro puntuaciones de los índices. Sin embargo, para no renunciar por completo a esas antiguas puntuaciones, en el Manual del WAIS-IV se señala que se puede usar el Índice de Comprensión verbal como CIV y el Índice de Razonamiento perceptual como CIE. Por otro lado, además del CIT, basado en los cuatro índices, el manual brinda una nueva puntuación llamada Índice de Capacidad General (ICG), que se basa sólo en las puntuaciones de Comprensión verbal y Razonamiento perceptual, con lo que se mantiene un paralelismo con el del CIT = CIV + CIE. Sólo el tiempo dirá qué será de esta mezcla de lo nuevo con lo viejo.

Puntuaciones y normas

La calificación del WAIS empieza determinando la puntuación natural de cada subprueba, la cual se convierte en puntuación estándar. El sistema de puntuaciones estándar de las subpruebas tiene una *M* = 10 y una *DE* = 3. (Regresa a la figura 3-10 del capítulo 3 para ver cómo se ajusta este sistema a la distribución normal.) En el WAIS, las puntuaciones estándar se denominan puntuaciones escalares. Hay diferentes cuadros de conversión de puntuaciones naturales en puntuaciones escalares para 13 grupos de edad, con intervalos que varían de 2 a 10 años (p. ej., 16-17, 20-24, 55-64, 85-89). También hay un grupo de referencia que incluye edades de 20 a 34 años.

Usar cuadros de conversión separados para cada grupo de edad significa que las puntuaciones escalares ocultan cualquier diferencia en el desempeño debida a la edad. En la edición previa, el procedimiento típico era convertir las puntuaciones naturales en puntuaciones escalares del grupo de referencia común, con lo que se preservaba cualquier cambio debido a la edad al representar el desempeño en puntuación escalar. Esto tiene consecuencias importantes para la interpretación de las puntuaciones. Hagamos una pausa para ilustrar este punto. El cuadro 8-7 muestra la conversión de la puntuación natural en escalar de dos subpruebas con base en el grupo de referencia (de 20 a 34 años de edad) y en el grupo de 85 a 90 años. Consideremos la puntuación natural de 24 en Diseño con cubos; ésta se convierte en puntuación escalar de 6 para el grupo de referencia, más de una *DE* abajo de la media o en un rango percentil de aproximadamente 10. Sin embargo, la misma puntuación natural de 24 se convierte en una escalar de 11 para el grupo de 85 a 90 años, que es una puntuación promedio. Así, la misma puntuación natural puede estar en el promedio para un grupo de normalización, mientras que para otro puede indicar un déficit considerable. Ahora consideremos un ejemplo de la subprueba Vocabulario. Tomemos la puntuación natural de 35; ésta se convierte en puntuación escalar de 10 tanto para el grupo de referencia como para el de 85 a 90 años de edad, es decir, no hay diferencia. Del mismo modo, una puntuación natural de 50 se convierte en una escalar de 15 en ambos grupos. Estos ejemplos ayudan a subrayar que la persona encargada de interpretar las puntuaciones debe comprender las características técnicas de la prueba; de lo contrario, podría hacer interpretaciones gravemente erróneas.

Cuadro 8-7. Conversiones de puntuaciones naturales en escalares de las subpruebas Vocabulario y Diseño con cubos del WAIS-IV para dos grupos de edad

Subprueba: / **Edad:**	**Vocabulario**		**Diseño con cubos**	
	20-34	85-90	20-34	85-90
Puntuación escalar				
−3 DE 1	0-2	0	0-5	0-2
2	3-5	1-2	6-8	3-4
3	6-9	3-5	9-13	5-6
−2 DE 4	10-13	6-9	14-18	7-8

5	14-17	10-13	19-23	9-10
6	18-21	14-17	24-28	11-12
−1 DE 7	22-25	18-22	29-33	13-14
8	26-29	23-26	34-38	15-17
9	30-33	27-30	39-43	18-20
Media 10	34-37	31-35	44-48	21-23
11	38-40	36-39	49-52	24-26
12	41-43	40-43	53-55	27-30
+1 DE 13	44-46	44-46	56-58	31-33
14	47-49	47-49	59-60	34-36
15	50-51	50-51	61-62	37-40
+2 DE 16	52-53	52-53	63	41-44
17	54	54	64	45-48
18	55	55	65	49-52
+3 DE 19	56-57	56-57	66	53-66

Fuente: *Manual de aplicación y calificación de WAIS®-IV*, Wechsler (2008b), cuadro 8-2.
Copyright © 2008 por NCS Pearson Inc. Reproducido con autorización. Todos los derechos reservados.

Las puntuaciones escalares de las subpruebas se pueden sumar y luego convertirse en puntuaciones compuestas, las cuales incluyen el CI Total y los cuatro índices (Comprensión verbal, Razonamiento perceptual, Memoria de trabajo y Velocidad de procesamiento). Todas éstas son puntuaciones estándar con M = 100 y DE = 15. El manual del WAIS proporciona los cuadros para convertir estas puntuaciones estándar en rangos percentiles; sin embargo, la mayor parte de los materiales de esta prueba emplean las puntuaciones estándar como base de la interpretación.

¡Inténtalo!

Usa el cuadro 8-7 para hacer las siguientes conversiones de puntuaciones naturales en escalares.

	Puntuación natural	**Puntuación escalar**	
		20-34 años	**85-90 años**
Vocabulario	46	_____	_____
Diseño con cubos	33	_____	_____

Supón que estas puntuaciones son del señor McInerney, de 85 años de edad. ¿Qué concluyes acerca del señor McInerney?

Estandarización «209-213a

WAIS se estandarizó con una muestra estratificada de 2450 adultos elegidos como representativos de la población de EUA de 16 a 89 años de edad. Las variables para la

estratificación incluyeron edad, sexo, raza/etnia, nivel educativo y región geográfica. Los grupos de edad básicos para elaborar las normas fueron 16-17, 18-19, 20-24, 25-29, 30-34, 35-44, 45-54, 55-64, 65-69, 70-74, 80-84 y 85-89. Cada grupo constó de 200 casos, a excepción de los cuatro grupos de mayor edad, donde el número fue menor. Los manuales del WAIS documentan con detenimiento la representatividad de los grupos de edad en términos de las variables de estratificación; también es encomiable que los manuales identifican los criterios para excluir ciertos tipos de casos. Por ejemplo, fueron excluidas las personas con demencia tipo Alzheimer, esquizofrenia, daltonismo, pérdida de audición o un trastorno de las extremidades superiores que afectan su desempeño motor. Así, lo mejor sería pensar en las normas como representantes de la población de adultos exentos de defectos sensoriales significativos y con una salud mental y física razonable.

Confiabilidad

Los manuales del WAIS-IV ofrecen un tratamiento sumamente minucioso de la confiabilidad. Informan la consistencia interna (división por mitades) y coeficientes de test-retest de CI, índices y subpruebas para cada grupo de edad por separado. También se informan los errores estándar de medición de todas las puntuaciones. Como señalamos, el uso del WAIS implica a menudo comparar varios índices y analizar el perfil de las subpruebas. Reconociendo este hecho, los manuales del WAIS brindan un tratamiento explícito de los errores estándar de las diferencias entre las puntuaciones, lo cual es muy loable.

La consistencia interna y la confiabilidad de test-retest del CI Total tienen un promedio de alrededor de .95 o superior, es decir, esta puntuación es confiable. El índice de Comprensión verbal (ICV) muestra un grado similar de confiabilidad, mientras que los otros tres (Razonamiento perceptual, Memoria de trabajo y Velocidad de procesamiento) tienden a tener confiabilidades un poco menores, con un promedio alrededor de .90, que sigue siendo un nivel alto de confiabilidad. Los errores estándar de medición de los cuatro índices tienen un promedio de 3 a 5 puntos escalares, mientras que en el CI Total el error estándar es de 2 puntos. Así, si alguien obtiene un CI Total de, digamos, 115 en el WAIS-IV, podemos estar bastante seguros de que esa persona obtendrá un CIT muy cercano a 115 si se le vuelve a aplicar la prueba una semana después, suponiendo que un examinador entrenado apropiadamente lleve a cabo la aplicación.

La confiabilidad de consistencia interna de las subpruebas del WAIS varía, por lo general, entre .70 y .90, con un promedio aproximado de .88. La confiabilidad de test-retest de las subpruebas tiene un promedio de .80; entre las subpruebas verbales, las confiabilidades de Vocabulario e Información son las más altas: por lo general, alrededor de .95 tanto en consistencia interna como en test-retest.

Validez

La información sobre la validez del WAIS es impresionante tanto por su amplitud como por su profundidad. Incluye miles de estudios en casi cualquier aspecto imaginable de la prueba. Los manuales proveen información sobre la validez de contenido, de criterio y de constructo. Como sucede con la mayoría de las pruebas de inteligencia, la discusión sobre la validez de contenido no es muy útil, ya que no hay un cuerpo de contenido bien definido al que podamos llamar "inteligencia". En cuanto a la validez de criterio, los manuales incluyen correlaciones con una gran variedad de pruebas, mientras que en lo referente a la validez de constructo, los manuales abordan la estructura factorial de la prueba y lo que éstos llaman estudios de comparación. La información sobre la estructura factorial, en general, apoya el uso de los cuatro índices, pues los reconoce como distintos. Los estudios de comparación muestran el patrón de las puntuaciones del WAIS de muchos grupos especiales, como de Alzheimer, Parkinson, problemas de aprendizaje y daño cerebral.

Interpretación del perfil

La interpretación del WAIS depende en gran parte del análisis del perfil de las puntuaciones, incluyendo las de las subpruebas y los índices. La figura 8-5 muestra la hoja del perfil tomada del Protocolo del WAIS-IV. En ella, hay un espacio para graficar las puntuaciones escalares de las subpruebas, así como los índices y el CIT. Otra página está diseñada para las "comparaciones de discrepancias" entre los cuatro índices, las diferencias entre las puntuaciones de las subpruebas y la media de todas las subpruebas, y las distintas "puntuaciones de proceso". Para todas estas comparaciones, en los espacios de resumen se puede registrar la significancia estadística de la diferencia y la frecuencia de ésta en las muestras de estandarización con base en los cuadros de datos de los manuales de la prueba. Estos resúmenes hacen hincapié en la necesidad de que el intérprete esté familiarizado con el concepto de errores de medición, que tratamos en el capítulo 4. Para un neófito, puede parecer que estos conceptos sólo son para impresionar; sin embargo, tienen un papel muy real en el trabajo cotidiano del psicólogo.

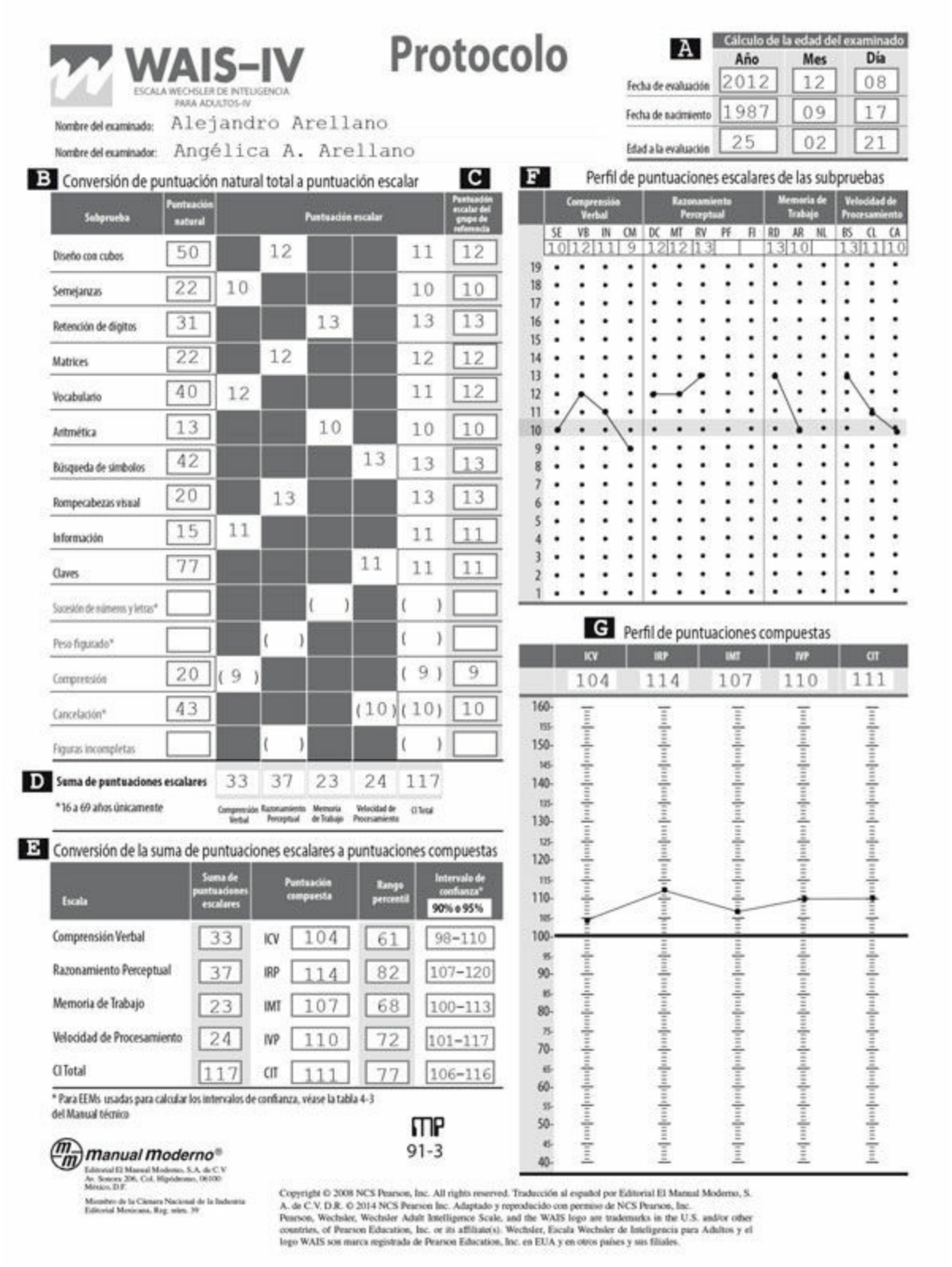


Figura 8-5. Hoja del perfil del protocolo del WAIS-IV.
*Fuente: Manual de aplicación y calificación de WAIS ®-IV (*p. 56). Copyright © 2008 por NCS Pearson, Inc. Reproducido con autorización. Todos los derechos reservados.

El manual recomienda seguir diversos pasos al interpretar el desempeño en el WAIS. El primer nivel se concentra en el CI Total, mientras que los siguientes niveles se ocupan de las comparaciones entre los índices. El manual recomienda examinar las discrepancias entre los índices y evaluar las fortalezas y debilidades. Los pasos finales, catalogados como "opcionales", tienen que ver con los patrones de las puntuaciones entre las subpruebas y el análisis de las "puntuaciones de proceso". Muchos sistemas interpretativos del WAIS y de pruebas similares emplean este tipo de método gradual, empezando con las puntuaciones más generales y siguiendo con las más específicas (lo cual, por cierto, corresponde más o menos a ir de las puntuaciones más confiables a las menos confiables).

Formas abreviadas

La aplicación del WAIS suele requerir de 60 a 90 min, que es mucho tiempo para el examinador y el examinado. Con los años, se han hecho numerosas propuestas de formas abreviadas o cortas del WAIS. En Groth-Marnat (2003) y House (1996) se puede encontrar un resumen de varias propuestas.

La editorial de las escalas Wechsler, Psychological Corporation, publicó el *Wechsler Abbreviated Scale of Intelligence* [Escala Abreviada Wechsler de Inteligencia], que ahora está en su segunda edición (WASI-II; Wechsler, 2011). En realidad se trata de dos versiones: una de cuatro subpruebas y otra de dos. El cuadro 8-8 muestra las subpruebas que se emplean en cada versión. La versión de dos subpruebas, que requiere de alrededor de 15 min para su aplicación, proporciona sólo el CI Total, mientras que la versión de cuatro subpruebas, que requiere cerca de 30 min para aplicarse, proporciona, además del CI Total, el índice de Comprensión verbal (ICV) y el de Razonamiento perceptual (IRP). El WASI tiene normas para edades de 6 a 90 años, de modo que abarca edades que el WISC y el WAIS cubren. Aunque se publicó por primera vez en 1999, WASI aún no ha aparecido en investigaciones sobre su uso como medida popular de inteligencia. Quizá lo sea en el futuro.

Cuadro 8-8. Subpruebas de la Escala Abreviada Wechsler de Inteligencia

Subprueba	Vocabulario	Semejanzas	Diseño con cubos	Matrices	Produce
Versión					
De cuatro subpruebas	x	x	x	x	ICV, IRP, CIT
De dos subpruebas	x			x	CIT

Escala Wechsler de Inteligencia para Niños – IV

La Escala Wechsler de Inteligencia para Niños – IV busca evaluar la capacidad intelectual de niños de 6-0 a 16-11 años. Hoy, WISC-IV es claramente la prueba de inteligencia de aplicación individual más usada con niños, pues desplazó al Stanford-Binet de esa posición de honor. Será práctico describir el WISC-IV comparándolo con WAIS-IV.

WISC frente a WAIS

Concebido originalmente como una extensión del WAIS, es comprensible que el WISC sea muy similar a él en su propósito y estructura. La mayor parte de lo dicho sobre el WAIS se puede aplicar al WISC; de hecho, hay muchas referencias cruzadas entre los manuales de estas dos pruebas. En el nivel más básico, la diferencia más evidente entre estas pruebas es el nivel de dificultad de los reactivos; por ejemplo, en la subprueba Vocabulario, en el WISC se podría preguntar cuál es el significado de "diccionario", mientras que en el WAIS se preguntaría cuál es el significado de "bibliografía". En Aritmética, el WISC preguntaría "¿Cuánto gastas si compras dos lápices en cinco pesos cada uno?", mientras que el WAIS preguntaría "¿Cuánto gastas si compras cuatro lápices en 12 pesos cada uno y dos libretas en 90 pesos cada una?". Desde luego, hay una superposición considerable en el nivel de dificultad entre los más difíciles del WISC y los más fáciles del WAIS.

Cuadro 8-9. Lista de subpruebas de los índices del WISC-IV

Índice	**Subpruebas principales**	**Subpruebas suplementarias**
Comprensión verbal	Semejanzas Comprensión Vocabulario	Información Palabras en contexto
Razonamiento perceptual	Diseño con cubos Matrices Conceptos con dibujos	Figuras incompletas
Memoria de trabajo	Retención de dígitos Sucesión de números y letras	Aritmética
Velocidad de procesamiento	Claves Búsqueda de símbolos	Registros
CI Total (CIT)	Suma de los cuatro índices	

En términos de estructura, el WISC y el WAIS son muy similares, pero hay algunas diferencias. El cuadro 8-9 enumera las subpruebas y las puntuaciones del WISC-IV. Si comparamos este cuadro con el 8-5 del WAIS-IV, podemos notar las siguientes

semejanzas:

- Las dos producen puntuaciones de CI Total, cuatro índices (también referidos en la métrica del CI) y de 10 a 15 subpruebas.
- La mayoría de las subpruebas son las mismas, aunque aparecen en un orden un poco diferente. También hay algunas diferencias en qué subpruebas son suplementarias. Las descripciones de las subpruebas del WAIS del cuadro 8-5 también son apropiadas para las del WISC; sólo hay que tener presentes las diferencias en los niveles de dificultad. Sin embargo, esta semejanza no necesariamente significa que las subpruebas miden los mismos rasgos en todos los niveles de edad y capacidad.

  Las *diferencias* entre el WISC-IV y el WAIS-IV son las siguientes:
- La composición de subpruebas de algunos índices es un poco diferente en las dos pruebas. Sin embargo, en general, los índices son muy parecidos en su composición.
- Un tipo de versión expandida del WISC-IV es el *WISC Integrated*, que ofrece pruebas adicionales, muchas con formato de opción múltiple, diseñadas para explorar varios procesos cognitivos, en especial de niños con discapacidades.

Características psicométricas del WISC-IV

El manual del WISC-IV ofrece una excelente presentación de las características psicométricas de la prueba. El programa de estandarización incluyó 2200 casos elegidos con mucho cuidado para que fueran representativos de la población de EUA de niños con edades de 6 a 16 años por género, región geográfica, raza/etnia y nivel educativo de los padres. La confiabilidad a nivel general y específico es muy similar a la del WAIS, que describimos antes. En términos de las medidas de consistencia interna, el CI Total (CIT) muestra de manera uniforme una $r = .97$, mientras que los índices varían, por lo general, de .90 a .95 y las subpruebas se ubican alrededor de .85. En cuanto a la estabilidad test-retest, el CIT tiene una confiabilidad promedio de .93, mientras que los índices rondan, por lo general, el .88 y las subpruebas, el .85. La confiabilidad interjueces de algunas subpruebas es prácticamente perfecta, pero incluso en el caso de las subpruebas que requieren del juicio al calificar, la confiabilidad es de, al menos, .95. En general, ésta es una prueba muy confiable, con la usual precaución de que mientras más corta es la prueba, menor es su confiabilidad. Los datos de validez incluyen correlaciones con un gran número de pruebas, análisis factoriales que apoyan las puntuaciones compuestas y descripciones del desempeño de numerosos grupos, como los de retraso mental y problemas de aprendizaje. En conjunto, los datos de la validez apoyan el uso del WISC para los propósitos para los que fue creado. Sin embargo, tenemos que repetir –como si fuera un mantra – que ninguna prueba de inteligencia es perfecta, por lo que se debe ser cauto siempre en la interpretación y buscar otras fuentes de información antes de sacar conclusiones acerca de una persona.

Stanford-Binet

Durante muchos años, la *Stanford-Binet Intelligence Scale* [Escala de Inteligencia Stanford-Binet] no tuvo rival como *la* medida de inteligencia humana. Hoy, en términos de la frecuencia con que se usa, ha sido eclipsada con claridad por las escalas Wechsler y compite con otras pruebas de capacidad mental de aplicación individual. Sin embargo, la Stanford-Binet aún se usa mucho en la práctica clínica; y lo más importante, conserva un lugar especial en la historia de la pruebas.

Recordemos del capítulo 1 que el francés Alfred Binet, en colaboración con Simon, creó las escalas Binet-Simon originales en 1905. Para su época, en que la mayoría de los teóricos se concentraban en los procesos perceptuales y sensoriales, fue revolucionaria la concepción de Binet de la inteligencia. Se centraba en lo que hoy llamaríamos "procesos mentales superiores", como el juicio y el razonamiento (véase Binet & Simon, 1916). En el cuadro 8-4 aparece la famosa definición de inteligencia que formuló Binet. La prueba de Binet no tuvo un nombre formal, sino que simplemente se refería a ella como escala. Las revisiones aparecieron en 1908 y 1911.

Varios estadounidenses prepararon versiones en inglés de las escalas Binet-Simon; la más famosa fue la que estuvo a cargo de Lewis Terman, que entonces trabajaba en la Universidad Stanford. Terman publicó su primera revisión en 1916; se trató de una revisión considerable de la obra de Binet. Lo que parecía distinguirla de otras revisiones estadounidenses fue la inclusión de una norma nacional desarrollada de manera muy avanzada para su tiempo. También introdujo un CI de razón, entidad destinada a penetrar en la conciencia nacional. Como Binet, Terman no tenía al principio un nombre formal para su prueba, pero pronto llegó a ser conocida como Stanford-Binet. En el discurso informal, la prueba simplemente se llama "Binet". La primera revisión de la prueba de 1916 apareció en 1937, con dos formas paralelas denominadas L y M, y un nuevo conjunto de normas nacionales. Una versión publicada en 1960 combinó las dos formas y se conoció como Forma L-M, pero no contó con una reestandarización; sin embargo, la revisión dejó fuera el CI de razón y optó por un CI de puntuación estándar. En 1970 apareció una reestandarización, pero con cambios mínimos en el contenido de la Forma L-M.

Transición a la nueva estructura: SB4 y SB5

En 1986 apareció una nueva edición, la cuarta, del Stanford-Binet (Thorndike, Hagen, & Sattler, 1986), y en 2003, una nueva revisión (SB5) (Roid, 2003a, 2003b). El cuadro 8-10 resume estos hitos en la historia del Stanford-Binet. Por lo común, una nueva edición de una prueba introduce algunos ajustes menores en la estructura y la actualización del contenido, las normas y los programas de investigación, pero esto no ocurrió con las ediciones más recientes del Stanford-Binet. En ellas se actualizaron el contenido y las normas, pero también se incluyeron desviaciones radicales respecto de las ediciones

previas en dos modos importantes. Primero, la escala clásica tipo Binet organizaba los reactivos de acuerdo con el nivel de edad; por ejemplo, el nivel de 6 años podía contener algunos reactivos de vocabulario, uno de retención de dígitos, algunos de semejanzas y un problema aritmético, todos ellos con un nivel de dificultad apropiado para esa edad. Segundo, la escala clásica tipo Binet producía una sola puntuación global; en la obra original de Binet, ésta era la edad mental, que evolucionó hasta llegar a ser el ahora familiar CI del Stanford-Binet.

Cuadro 8-10. Hitos en el desarrollo del Stanford-Binet	
1905, 1908, 1911	Ediciones de las escalas originales Binet-Simon
1916	Revisión en Stanford de las escalas Binet; se empleó el CI de razón
1937	Formas L y M, nueva estandarización
1960	Una forma (L-M), sin reestandarización, CI de puntuación estándar
1972	Nueva estandarización
1986	Cuarta edición (SB4), puntuaciones múltiples, nueva estandarización
2003	Quinta edición (SB5), más puntuaciones, nueva estandarización, cambios en la DE del CI

SB4 y SB5 renunciaron a esas dos características. Aunque conservan algunos reactivos de sus predecesores inmediatos, en esencia, son pruebas nuevas en su totalidad más que la evolución a partir de la edición previa. En SB4 y SB5, los reactivos están organizados por subprueba, siguiendo la tradición de las escalas Wechsler. También presentan puntuaciones múltiples además de la puntuación total. Es importante reconocer estos cambios estructurales, porque muchas referencias al "Binet" o al "Stanford-Binet" describen la antigua estructura tradicional más que la prueba actual.

Cuadro 8-11. Organización del Stanford-Binet, Quinta Edición

Factores	Dominios	
	No Verbal (NV)	**Verbal (V)**
Razonamiento fluido (RF)	***Razonamiento fluido no verbal**** Actividades: Series de objetos/ Matrices (iniciación)	***Razonamiento fluido verbal*** Actividades: Razonamiento temprano (2-3), Disparates verbales (4), Analogías verbales (5-6)
Conocimiento (CO)	***Conocimiento no verbal*** Actividades: Conocimiento procedimental (2-3), Disparates en imágenes (4-6)	***Conocimiento verbal**** Actividades: Vocabulario (iniciación)
Razonamiento cuantitativo (RC)	***Razonamiento cuantitativo no verbal*** Actividades: Razonamiento cuantitativo (2-6)	***Razonamiento cuantitativo verbal*** Actividades: Razonamiento cuantitativo (2-6)
Procesamiento viso-espacial (VE)	***Procesamiento viso-espacial no verbal*** Actividades: Tablero de formas (1-2), Patrones de formas (3-6)	***Procesamiento viso-espacial verbal*** Actividades: Posición y dirección (2-6)
Memoria de trabajo (MT)	***Memoria de trabajo no verbal*** Actividades: Respuesta demorada (1), Retención de bloques (2-6)	***Memoria de trabajo verbal*** Actividades: Memoria de enunciados (2-3), Última palabra (4-6)

Nota: Los nombres de las 10 subpruebas aparecen en ***cursivas negritas***. Las actividades se muestran con el nivel en que aparecen. *Subpruebas de encaminamiento

Adaptado de *Stanford-Binet Intelligence Scales (SB5), Fifth Edition.* Por Gale H. Royd, 2003, Austin, Tx: PRO-ED., Inc, copyright 2003 por PRO-ED, Adaptado con autorización. Todos los derechos reservados.

El cuadro 8-11 esboza la estructura de SB5. Recordemos del capítulo 7 nuestra discusión de los modelos jerárquicos de la inteligencia, en especial el de Carroll. El manual de SB5 (Roid, 2003a), de un modo bastante explícito, adopta este modelo aunque reconoce que el SB5 no cubre todas las facetas del modelo. El cuadro 8-11 debe verse como una matriz: renglones × columnas. La matriz revela los tipos de puntuación que produce SB5. En el nivel superior se encuentra el CI Total (CIT), muy parecido al de las escalas Wechsler; al igual que en estas escalas, el CIT significa "*g*". Como lo sugiere el resumen, SB5 también produce otros dos tipos de puntuaciones compuestas. El primero incluye el CI Verbal y el No Verbal, que resultan de sumar el contenido de las dos columnas del cuadro. El segundo tipo incluye cinco índices:

- Razonamiento fluido
- Conocimiento
- Razonamiento cuantitativo

- Procesamiento viso-espacial
- Memoria de trabajo

Las puntuaciones de estos índices resultan de la suma del contenido de los renglones del cuadro. Es decir, cada una tiene un componente verbal y no verbal. Los índices no se denominan CI, pero utilizan la métrica del CI familiar: M = 100, DE = 15. En este aspecto, podemos notar que SB5 ha claudicado al adoptar DE = 15, que es la tradición de las escalas Wechsler, después de haber usado por tantos años DE = 16. Por último, igual que las escalas Wechsler, SB5 proporciona puntuaciones para cada celda de la matriz: un total de 10, las cuales tienen M = 10 y DE = 3.

Una característica importante del SB5 es el uso de "pruebas de iniciación", las cuales están señaladas en el cuadro (*). Se trata de las subpruebas Series de objetos/Matrices en el dominio No Verbal y de Vocabulario en el Verbal. Todas las pruebas están organizadas en niveles, cinco no verbales y seis verbales; el examinador aplica primero estas pruebas de iniciación y usa los resultados para determinar el nivel apropiado de las otras subpruebas.

Otra característica que es muy importante en el SB5, y que no es clara en el cuadro 8-11, es el rango de edades excepcionalmente amplio que abarca: de 2 a 85 y más años. Así, mientras que Wechsler ofrece diferentes pruebas (WPPSI, WISC y WAIS, similares, pero no idénticas en su estructura) a lo largo de este rango de edades, el SB5 cubre todo este rango. Es evidente que el nivel de dificultad de los reactivos varía a lo largo de cada rango, pero también hay algunos cambios en el tipo de reactivos.

Características psicométricas del SB5

En general, las características psicométricas –estandarización, confiabilidad, validez– del SB5 son excelentes. El conjunto de datos es comparable con el que presentamos en el caso de las escalas Wechsler. Los fundamentos de su estructura se presentan de manera cuidadosa. Las normas se basan en muestras elegidas con cuidado y meticulosidad de todo el rango de edad de 2 a 85y más años. Se hicieron grandes esfuerzos para asegurar la neutralidad, es decir, la ausencia de sesgos. Las confiabilidades de consistencia interna, por lo general, están arriba de .90 en el caso de las diversas puntuaciones totales: Total, Verbal, No Verbal e Índices. Los coeficientes de confiabilidad test-retest son ligeramente inferiores. Como siempre sucede, la confiabilidad de las subpruebas es inferior, pero, en general, se ubican dentro de un rango respetable. Por último, como en el caso de las escalas Wechsler, el manual del SB5 presenta una gran cantidad de información relacionada con la validez: correlaciones con ediciones anteriores y con otras pruebas, así como diferencias de grupos como intelectualmente sobredotados, retraso mental y problemas de aprendizaje. En conjunto, el manual del SB5 ofrece un modelo de presentación profesional. La eterna pregunta es: ¿cuál es mejor, el Wechsler o el Stanford-Binet? Dejamos esta pregunta abierta, pero una cosa sí es segura: ahora son mucho más parecidas que en el pasado.

¡Inténtalo!

Compara la estructura del SB5 que aparece en el cuadro 8-11 con el modelo jerárquico de Carroll que se presentó en la página 184a». ¿Qué partes del modelo de Carroll encuentras? ¿Cuáles no aparecen?

Pruebas breves de capacidad mental de aplicación individual

Entre las pruebas de capacidad mental de aplicación individual, las más usadas requieren de 60 a 90 min e incluyen varios tipos de reactivos. Sin embargo, también hay pruebas de aplicación individual que se caracterizan por su brevedad y sencillez, y que se emplean cuando el clínico necesita una evaluación rápida y global de la capacidad mental. A veces, el esfuerzo que requiere una evaluación más extensa y detallada no está justificado. El trabajo subsiguiente con el cliente puede sugerir que es necesaria una evaluación más detallada de la capacidad mental; sin embargo, una evaluación rápida puede ser suficiente al inicio.

Las pruebas de la categoría "breve" suelen requerir menos de 15 min para su aplicación, y a veces sólo 7 u 8 min. Por lo común, estas pruebas proporcionan una sola puntuación global y, a menudo, constan de un solo tipo de reactivos. La información sobre la validez de ellas se concentra en mostrar que tienen una correlación razonablemente alta con medidas más extensas de capacidad mental.

Prueba de Vocabulario en Imágenes Peabody

El **Peabody Picture Vocabulary Test** [Prueba de Vocabulario en Imágenes Peabody], Cuarta Edición (**PPVT**-4) ofrece un excelente ejemplo de una prueba sencilla y breve de capacidad mental. Requiere de 10 a 15 min para su aplicación. Están disponibles las normas para edades de dos años y medio (2 1/2) hasta 90 y más años, un rango excepcionalmente amplio para un solo instrumento. La prueba consta de 228 reactivos; en cada uno, el examinador lee una sola palabra y el examinado elige, entre cuatro imágenes, la que la represente mejor. Así, aunque es de aplicación individual, la prueba emplea reactivos con formato de opción múltiple. La figura 8-6 presenta ejemplos ficticios de reactivos. El examinador podría decir "Muéstrame la *pelota*". Para ilustrar de qué manera un reactivo mucho más difícil puede usar el mismo formato, el examinador podría decir "Muéstrame *esfera*". El PPVT ha aumentado su popularidad de manera constante desde su primera publicación en 1959. La más reciente edición, la que describimos aquí, data de 2007. Varias investigaciones muestran que el PPVT ahora supera en uso al Stanford-Binet, pero aún está muy lejos del WAIS y WISC.

Figura 8-6. Reactivos ficticios del formato del PPVT (no son reactivos reales).

Propósitos

De acuerdo con el manual, el PPVT-4 evalúa "la comprensión de las palabras habladas en inglés estándar y, así, es una medida del aprovechamiento del examinado en la adquisición de vocabulario" (Dunn & Dunn, 2007, p. 2). Al parecer, el uso típico del PPVT es como una prueba corta de capacidad mental. La clara mayoría de los estudios de validez resumidos de la literatura profesional, como se informa en las Referencias técnicas (Williams & Wang, 1997), tratan la relación entre el PPVT y medidas más extensas de inteligencia como el Stanford-Binet y las escalas Wechsler. Nadie trata de

usar el PPVT como sustituto del Stanford Achievement Test [Prueba de Aprovechamiento de Stanford], pero, según parece, muchas personas tratan de usarlo como sustituto del Stanford-Binet. El PPVT aprovecha el hecho de que el conocimiento de palabras, sea en forma de vocabulario expresivo o receptivo, tiene una correlación alta con medidas de inteligencia más extensas y complicadas. La razón de que el vocabulario sea tan buena medida de "*g*" no está del todo clara; sin embargo, muchos estudios muestran una alta correlación entre el simple conocimiento de palabras y otras medidas de inteligencia.

Materiales y aplicación

Mientras que en las pruebas integrales de capacidad mental y aplicación individual, como las escalas Wechsler, aparecen complejos conjuntos de materiales, el PPVT es la esencia de la sencillez. Un solo caballete de aplicación contiene todos los reactivos; cada uno aparece en una página dentro del caballete. El PPVT proporciona un solo manual de 209 páginas que incluye información sobre estandarización, confiabilidad, validez y normas. El aplicador registra las respuestas en un sencillo cuadernillo de ocho páginas, que ordena los reactivos en sus 19 bloques (véase más adelante).

Los arreglos físicos para la aplicación del PPVT son, en esencia, los mismos que los de pruebas más extensas. El examinador se sienta con el examinado frente a una mesa y tiene que establecer el *rapport*. Hay reactivos muestra para introducir la tarea. El examinado mira las imágenes en la página del equipo de la prueba. El examinador dice una palabra y el examinado emite su respuesta señalando o diciendo el número de la imagen.

El examinador debe elegir un punto de inicio y, luego, establecer un nivel basal y de tope. Los 228 reactivos están ordenados en 19 bloques de 12 reactivos cada uno; aumentan gradualmente de dificultad, desde las palabras que deberían estar en el vocabulario de trabajo de la mayoría de niños preescolares hasta las que se considerarían propias de una tesis de maestría de un intelectual. Esta organización es evidente para el examinador, pero no para el examinado. El examinador elige el reactivo de inicio, el primero de uno de los 19 bloques, con base en la edad del examinado; si hay una buena razón para suponer que el examinado está por arriba o por debajo del promedio para su edad, el examinador puede elegir un reactivo de inicio superior o inferior. El examinador establece un conjunto basal, el bloque de 12 reactivos más bajo en que el examinado no comete más de un error. El examinado recibe crédito por todos los reactivos que se encuentran por debajo del conjunto basal. La aplicación continúa hasta llegar al conjunto tope, el cual es el bloque de 12 reactivos más alto en que, al menos, ocho de las respuestas son incorrectas. Después de alcanzar el conjunto tope, se detiene la prueba. De acuerdo con el manual, por medio de este procedimiento, un individuo típico responde cinco bloques, es decir, 60 reactivos.

Nuestra descripción del procedimiento de aplicación del PPVT ilustra un punto importante. Incluso tratándose de una prueba tan sencilla como ésta, el procedimiento de

aplicación puede ser muy detallado, por lo que es esencial que se siga al pie de la letra para que la aplicación se considere estandarizada. La violación del procedimiento puede hacer que las normas sean inaplicables; por ejemplo, supongamos que el examinador siguió aplicando reactivos después de alcanzar el conjunto tope. Sin duda, el examinado tendrá algunos reactivos correctos por el azar, lo que añada, quizá, 5 o 6 puntos a su puntuación natural, lo cual podría llevar a sobreestimar gravemente la capacidad de la persona.

Puntuaciones, normas y estandarización

Dentro del conjunto tope, el examinador toma nota del número del reactivo más alto, llamado reactivo tope. La puntuación natural de la prueba es el número de este reactivo menos el número de respuestas incorrectas; después, ésta se convierte en una puntuación estándar normalizada y/o equivalente de edad. El sistema de puntuación estándar tiene M = 100 y DE = 15, el mismo sistema que el CI de las escalas Wechsler. El manual del PPVT nunca se refiere a esta puntuación estándar como CI, pero uno tendría que preguntarse por qué habría de usarse este sistema específico de puntuaciones estándar si no es para insinuar un CI. Se cuenta con sistemas separados de puntuaciones estándar para cada grupo de edad que abarcan intervalos variables para cubrir el rango de 2:6 a 90:11. Las puntuaciones estándar pueden convertirse en rangos percentiles, estaninas y equivalentes de curva normal por medio de un cuadro similar al cuadro 3-1.

Los equivalentes de edad, basados en la mediana del desempeño de grupos de edad sucesivos, varían entre "< 2 : 0" (menos de dos años, cero meses) y "> 24 : 11" (más de 24 años, 11 meses). Recordemos que los equivalentes de edad carecen de sentido cuando el rasgo que medimos deja de aumentar (véase Capítulo 3, Fortalezas y debilidades de las normas de desarrollo). La figura 8-7 proporciona un ejemplo claro de este fenómeno; en ella se grafica la mediana de las puntuaciones naturales por grupos de edad de la estandarización del PPVT. Podemos notar que la curva es muy pronunciada de los 2 a cerca de los 15 años de edad; luego se eleva lentamente hasta casi los 25 años; después, casi es plana durante toda la etapa adulta, con una ligera disminución en los grupos de edad mayores.

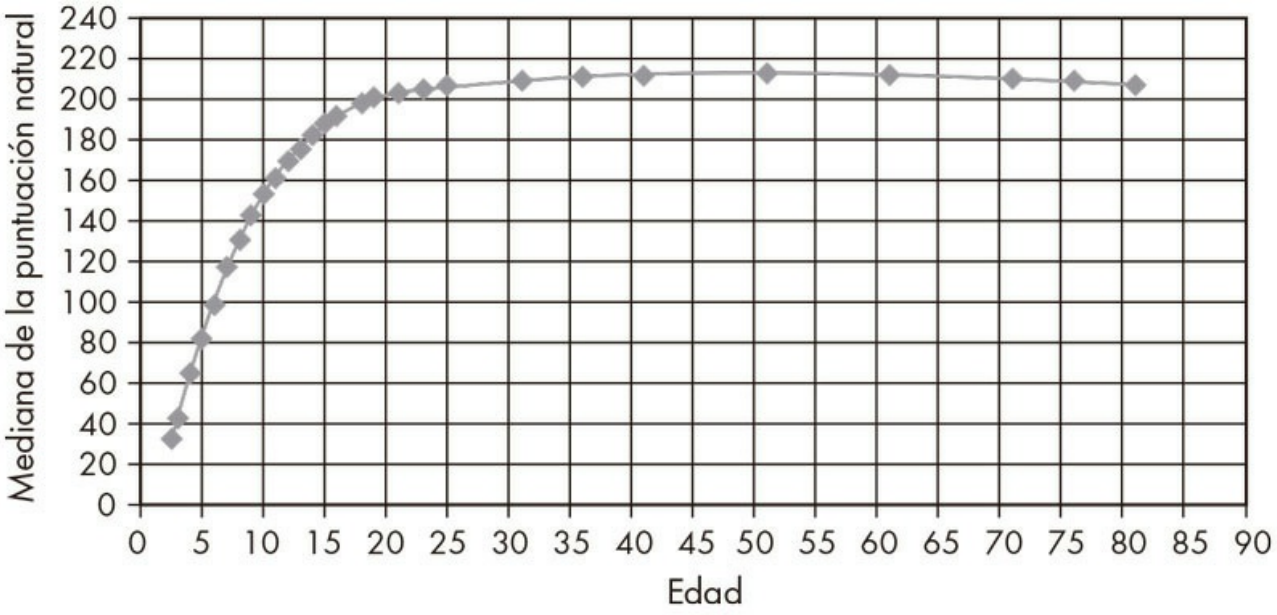


Figura 8-7. Crecimiento en vocabulario por edad en el PPVT-4.

¡Inténtalo!

Utiliza lo que sabes acerca de determinar los equivalentes de edad con la figura 8-8. ¿Qué equivalente de edad corresponde a una puntuación natural de 130?

Las normas del PPVT se basan en muestras de 60 a 200 casos por cada grupo de edad; son 28 grupos de edad que van de 2:6-2:7 a 81-90:11, y están conformados en total por 3540 casos. Éstos se eligieron buscando que fueran representativos de la población de EUA en términos de edad, género, región geográfica, raza/etnia y estatus socioeconómico, tomando como indicador de este último el nivel educativo. El manual documenta de manera cuidadosa la representatividad de las muestras de estandarización respecto de estas variables. Los examinadores excluyeron casos en que había problemas de visión o audición, o un dominio limitado del inglés. El manual no indica cuántos de estos casos se excluyeron. Como sucede con muchos proyectos de estandarización, los examinadores reclutaron a los participantes, y no es fácil determinar qué efectos puede tener el proceso de reclutamiento.

Confiabilidad

El manual del PPVT proporciona un excelente conjunto de datos sobre la confiabilidad, que incluyen dos tipos de consistencia interna (división por mitades y alpha), formas alternas y test-retest. Los coeficientes de confiabilidad de consistencia interna se basan en todos los casos del programa de estandarización que describimos antes. Una submuestra

de 508 casos de la muestra de estandarización respondió ambas formas de la prueba, en lo cual se basan los datos de confiabilidad de formas alternas. Los datos de test-retest provienen de submuestras del grupo de estandarización a las que se volvió a aplicar la prueba un mes después aproximadamente. La confiabilidad de consistencia interna y de formas alternas en varios grupos de edad está alrededor de .95, con medias que van de .94 a .97. La confiabilidad de formas alternas varía entre .87 y .93 en los distintos grupos de edad, con una media de .89. La confiabilidad de test-retest varía de .92 a .96, con una media de .93. En general, los datos de la confiabilidad del PPVT-4 son extensos y favorables.

Los datos de la confiabilidad de esta prueba ilustran un principio general importante en el campo de las pruebas psicológicas. Tratándose de un rasgo con una definición restringida (conocimiento del vocabulario), un número adecuado de reactivos (un promedio de 60 por examinado) y procedimientos de elaboración de la prueba minuciosos, podemos obtener una medición muy estable, incluso con un tiempo de aplicación relativamente breve.

Validez

El manual del PPVT aborda los temas de validez de contenido, de constructo y de criterio; esta última es la que recibe, por mucho, la mayor atención. La mayoría de los estudios sobre la validez de criterio informan correlaciones entre el PPVT y otras pruebas de capacidad mental. En el manual también se resumen estudios con poblaciones especiales, como individuos con retraso mental, problemas de aprendizaje o intelectualmente sobredotados. Muchos estudios informan las correlaciones con el WISC, el SB y pruebas integrales completas de capacidad mental. Este patrón sugiere que, al menos en la literatura publicada, el principal interés es el uso del PPVT como un sustituto rápido de las medidas más extensas de inteligencia.

¿Qué tan altas son las correlaciones del PPVT con las medidas más extensas de inteligencia? En promedio, las correlaciones son muy altas; es comprensible que éstas sean especialmente altas con medidas más verbales, como las puntuaciones verbales de las escalas Wechsler, mientras que con las medidas no verbales las correlaciones son más bajas. Aunque los resultados varían a lo largo de docenas de estudios sobre la validez de criterio, no es inusual encontrar correlaciones del PPVT de .80 con medidas como el CIV Wechsler o el compuesto Stanford-Binet.

El PPVT ilustra que, al menos, para ciertos propósitos muy sencillos, el instrumento proporciona información confiable y útil. De hecho, en realidad parece muy destacable que, en cuestión de unos 15 min, se pueda obtener una aproximación bastante buena a la información que brinda la puntuación total de una prueba como WISC. Desde luego, el PPVT no permite un análisis del perfil o de la discrepancia entre las puntuaciones, pero, a veces, no se necesitan todos esos detalles.

Hay una lección final que aprender del PPVT. Los autores de la prueba, así como los usuarios, ansían sacar más información de una prueba de la que ésta puede dar. La

cuarta edición del PPVT ilustra este punto. Mientras que todas las ediciones previas de la prueba se conformaban con proporcionar una sola puntuación general, con información amplia acerca de la confiabilidad y validez de esa puntuación, el PPVT-4 recomienda contar e intentar interpretar puntuaciones separadas de los reactivos basados en sustantivos, verbos y "atributos" (es decir, adjetivos y adverbios). El manual no ofrece un fundamento para tales interpretaciones, ni normas de las puntuaciones, ni información acerca de la confiabilidad o validez de dichas puntuaciones. La lección para el lector es: hay que estar alerta a qué usos de la prueba pueden tener un adecuado fundamento técnico y renunciar a los usos que carecen de él.

Dos entradas más

Para ilustrar las pruebas individuales de inteligencia, hemos ofrecido descripciones del WAIS, WISC, SB y PPVT; hay muchas otras pruebas de inteligencia de aplicación individual, demasiadas para enumerarlas aquí. Sin embargo, otras dos merecen ser mencionadas, porque se citan con frecuencia en la literatura de las pruebas psicológicas. Se trata del *Woodcock-Johnson Tests of Cognitive Abilities* [Pruebas Woodcock de Capacidades Cognitivas], Tercera Edición (WJ-III), y el *Kaufman Assessment Battery for Children* [Batería Kaufman de Evaluación para Niños] (K-ABC). Animamos al lector a consultar las fuentes usuales para obtener más información acerca de estas dos pruebas, empezando por los sitios de internet de las editoriales, que aparecen en el apéndice C.

Una prueba de capacidad mental específica: Escala de Memoria de Wechsler

Las pruebas como el WAIS, WISC y SB intentan medir el funcionamiento intelectual general por medio de una amalgama de contenido que cubre áreas como significado verbal, razonamiento cuantitativo, relaciones espaciales, agudeza perceptual, memoria y otras tareas intelectuales. Son como una canasta de mercado: un bote de esto, una bolsa de aquello, una docena de éstos, dos kilos de aquello. A veces, el psicólogo quiere medir un área con mayor profundidad. La capacidad específica que recibe más atención es la memoria; ésta es crucial para el aprendizaje, que, a su vez, es la base del futuro desarrollo intelectual. La memoria parece tener una sensibilidad particular a los cambios en el funcionamiento cerebral, los cuales pueden ser resultado de un trauma en la cabeza, el envejecimiento u otros padecimientos. La figura 8-8 describe las relaciones entre la medición de la capacidad mental general y la capacidad mental específica de la memoria. Una medida de capacidad mental general consta de diversas subáreas. Sin tratar de dar una lista exhaustiva, incluimos vocabulario (V), relaciones verbales (RV), razonamiento cuantitativo (RC), relaciones espaciales (E), capacidad perceptual (P) y memoria (M). Existen otras subáreas que no incluimos, por lo que nuestro catálogo no está completo. Además, la memoria misma puede ser dividida en varias áreas específicas, como se

indica en la figura 8-8. Identificaremos algunas de ellas al describir la siguiente prueba.

Entre las muchas pruebas que buscan medir de manera específica la memoria, sin duda, la más usada es la *Wechsler Memory Scale* [Escala de Memoria de Wechsler], ahora en su cuarta edición (WMS-IV, Wechsler, 2009a, 2009b). Esta prueba no sólo es la más usada para medir la memoria, sino que también es la más usada de todo el repertorio del psicólogo. Siempre se ubica entre las 10 pruebas más usadas en cualquier investigación del uso de pruebas, por lo que presentamos aquí una breve descripción de este instrumento.

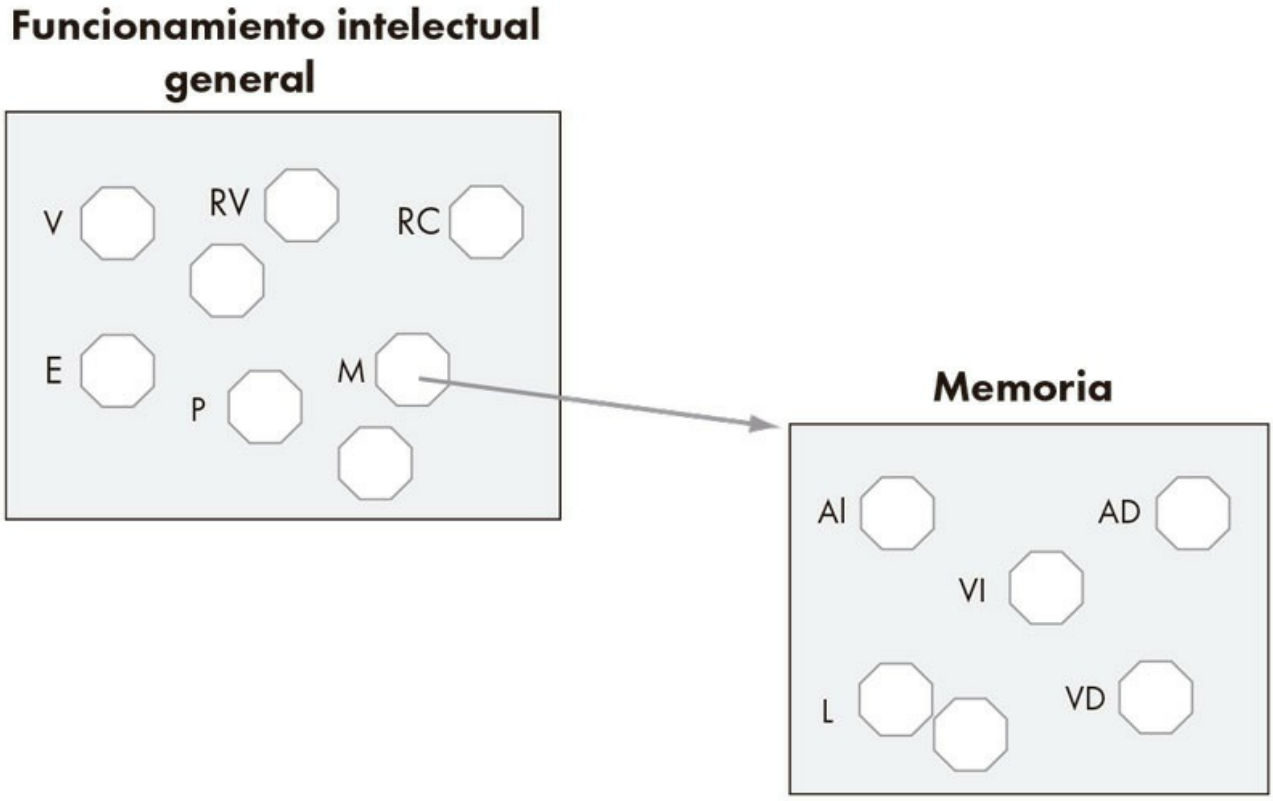


Figura 8-8. Relación entre una medida de funcionamiento mental general y una de memoria.

Estructura y reactivos

Como se mencionó en los párrafos anteriores, el WMS busca medir la memoria de una manera clínicamente pertinente abordando diversas funciones de memoria. La prueba está diseñada para edades de 16 a 90 años; la edición actual, por primera vez, cuenta con dos baterías: la de adultos, para edades de 16 a 69, y la de adultos mayores, para edades de 69 a 90 años. Esta última simplemente omite ciertas subpruebas de la forma para adultos. Aquí describiremos esta última batería.

La aplicación del WMS-IV es similar a la de las pruebas individuales de inteligencia, es

decir, hay un examinador que establece el *rapport* y presenta los estímulos materiales al examinado, que está sentado frente a una mesa. El examinador califica los reactivos de inmediato, emplea reglas de inicio y discontinuación, y registra sus observaciones conductuales. El tiempo de aplicación, igual que con otras pruebas de inteligencia, varía en cada caso, pero suele requerir cerca de 80 min.

El WMS es primordialmente una medida de memoria a corto plazo. Las tareas de memoria varían desde el recuerdo inmediato hasta un periodo de 30 min. Esta prueba no intenta medir la memoria a largo plazo más que de una manera muy indirecta, por ejemplo, lo que un individuo recuerda de algún curso al que haya asistido un año antes. ¿Cómo podríamos medir la memoria a corto plazo? Pensemos en algunas de las tareas sencillas que encontramos a diario que hacen uso de la memoria. También pensemos en las tareas de memoria cuya demostración hemos presenciado en el laboratorio de psicología. Ahora tomemos todas estas tareas y lancémoslas en forma de reactivos, con instrucciones estandarizadas, criterios muy específicos de calificación y un cuadernillo para registrar las respuestas; desde luego, agreguemos normas, estudios de confiabilidad y demostraciones de validez. Eso es, en esencia, lo que hace el WMS. Observaremos ejemplos de tareas específicas de la memoria en la siguiente sección.

El cuadro 8-12 bosqueja la estructura del WMS-IV. Está conformado por 10 subpruebas que producen cuatro puntuaciones de índice. No es evidente en qué consisten las tareas si vemos sus nombres, así que daremos algunos ejemplos. No intentaremos dar descripciones detalladas de todas las subpruebas, sino sólo lo suficiente para formarnos una idea de la prueba.

Cuadro 8-12. Estructura y subpruebas del WMS-IV

Dominio	**Subpruebas primarias**	**Subpruebas opcionales**
Auditivo/Verbal	Memoria lógica I y II Pares verbales asociados I y II	Listas de palabras I y II
Visual/No Verbal	Imágenes de familia I y II Caras I y II	Reproducción visual I y II
Memoria de trabajo	Sucesión de números y letras Retención espacial	Control mental
Exploración general		Información y orientación

El WMS usa dos distinciones importantes. La primera contrasta la entrada de información auditiva y visual; en algunos reactivos, el examinado escucha los estímulos y, en otros, los observa, ya que la información más importante en la vida cotidiana proviene de estos dos sentidos. La prueba no recurre a otras modalidades sensoriales. La segunda distinción es entre la **memoria inmediata** y **demorada**.

En el WMS, memoria inmediata significa recordar pocos segundos después de la presentación del estímulo; no significa recordar en milisegundos, como en un estudio de imágenes remanentes o de fenómenos de ese tipo. Memoria demorada significa recordar de 20 a 30 min después de la presentación del estímulo, dentro del mismo periodo de la

aplicación; no significa recordar después de días, meses o años. La figura 8-9 bosqueja estas dos distinciones; en muchas de las subpruebas del WMS sólo se llenan los cuadros que aparecen en esta figura. Podemos notar que la entrada de estímulos auditivos implica material verbal, mientras que la de estímulos visuales, material no verbal.

	Periodo del recuerdo	
Tipo de estímulo	Inmediato	Demorado
Auditivo		
Visual		

Figura 8-9. Dos distinciones importantes en el WMS: tipo de estímulo y periodo del recuerdo.

Al describir algunas subpruebas, podemos notar el contraste con los reactivos de pruebas de capacidad mental general. No hay dificultad conceptual con los reactivos de memoria: no hay palabras difíciles como una prueba de vocabulario, ni abstracciones como en una prueba de semejanzas o analogías, ni se requieren manipulaciones complicadas de material cuantitativo. A simple vista, los reactivos de memoria son bastante sencillos. La pregunta es: ¿puedes recordar este simple material?

El nombre **Memoria lógica** es un poco inadecuado, pues no implica ninguna lógica. El estímulo se presenta simplemente en un párrafo lleno de detalles con cerca de 50 palabras que se lee al examinado, a quien se pide que repita la historia al examinador. Éste registra el número de detalles que el examinado repitió. Por ejemplo, el examinador lee el pasaje que aparece en la figura 8-10; cuando termina, el examinado repite todo lo que pueda recordar. El manual de la prueba proporciona los criterios que debe cumplir la respuesta para ser considerada aceptable. También hay una versión "demorada" de esta tarea, en la que se pide al examinado que repita la historia 30 min después. Las versiones demoradas corresponden a "II" en el cuadro 8-12.

Párrafo de memoria lógica

El examinador lee:

Ned y Abigail viajaron en tren de Nueva York a Washington. Aunque este viaje suele durar sólo 3 hr, hubo un retraso por una tormenta de nieve, así que no llegaron sino casi a medianoche. Ned vivía al este de la estación de tren y Abigail, al oeste. Sólo había un taxi, así que ambos lo tuvieron que abordar; primero llevó a Abigail, y luego a Ned en la dirección opuesta.

Luego se pide al examinado que repita la historia mientras el examinador registra cuántos elementos de la historia están incluidos en la respuesta.

Asociación de pares de palabras

El examinador lee una vez estos pares.

lápiz – abajo	carro – bebé	cocer – cuatro	pie – verde
montaña – letra	papel – piso	árbol – abierto	bloque – toalla

Se pide al examinado que diga la palabra correcta que complete un par cuando el examinador lee una de ellas, por ejemplo, "cocer", luego "árbol", etc. Las palabras estímulo no siguen el orden de la presentación original. Después, el examinador dice los pares por segunda vez y continúa con las palabras estímulo. El ciclo se repite varias veces más.

Reproducción visual

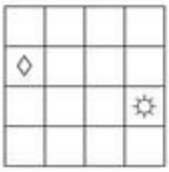
Se muestra al examinado una figura durante 10 segundos, luego se quita y se le pide que la reproduzca. Aquí hay una figura muestra.

Control mental

Contar de cuatro en cuatro en orden descendente empezando con 37.

Diseños

Se muestra al examinado esta cuadrícula con un diseño insertado durante 10 segundos.



Se presenta una nueva cuadrícula y un conjunto de cartas con diseños. Se pide al examinado que inserte las tarjetas en los lugares correctos.

Retención de símbolos

Retención de símbolos funciona exactamente como Retención de dígitos (véase cuadro 8-2), excepto porque usa letras en vez de números, por ejemplo, G-M-B-Y-T.

Figura 8-10. Ejemplos de reactivos de memoria similares a los del WMS.

Observemos el reactivo de Reproducción visual de la figura 8-10; es un reactivo sencillo. Desde luego, se pueden presentar figuras mucho más complicadas. Estos reactivos, como los de Memoria lógica y Asociación de pares, también pueden aplicarse en el tipo de memoria "demorada".

Las puntuaciones de las subpruebas y los índices, descritas antes, proporcionan el principal marco interpretativo del WMS. Sin embargo, la prueba también incluye puntuaciones de *proceso* derivadas de las subpruebas. Además, el Protocolo y el Manual de aplicación y calificación del WMS dedican considerable atención a informar numerosas "puntuaciones de contraste" y "puntuaciones de diferencia". Por ejemplo, el Protocolo demanda el contraste entre el índice de Memoria auditiva y el de Memoria

visual, y entre el Índice de Capacidad General del WAIS y el índice de Memoria de trabajo visual del WMS, así como la diferencia entre el índice de Memoria demorada y cada una de las subpruebas que lo componen. Las puntuaciones "de contraste" no corresponden a diferencias simples, sino que se basan en normas separadas por subgrupos en una escala. En conjunto, el Protocolo demanda un conjunto desconcertante de información, lo cual es consistente con la tendencia que hemos notado en todos lados: sacar más y más puntuaciones de una prueba. La persona que interpreta todas estas puntuaciones necesita tener muy presente que mientras más combinaciones de puntuaciones hay, también hay más oportunidades de incluir errores de medición (y, por lo tanto, de sacar conclusiones injustificadas).

Extrañamente, el WMS no produce una puntuación total a diferencia del WAIS y WISC que sí lo hacen: el CI Total. Parece sensato pensar en una puntuación total de memoria y, de hecho, el manual técnico del WMS (Wechsler, 2009b, p. 62) señala que "todos los factores [en el análisis factorial del WMS] tienen correlaciones muy altas...". Como una evidencia más, la correlación entre los índices de Memoria inmediata y Memoria demorada es de .87 en la muestra de estandarización. Toda esta información parece estar a favor de una puntuación total, pero el WMS no la proporciona.

La edición previa incluía una serie de cuestiones, a las que se refería como *Información y orientación*, para determinar el estatus mental grueso de una persona, por ejemplo, saber qué día de la semana es y recordar una lista de unas pocas palabras. El WMS-IV eleva estos reactivos, 12 en total, a una subprueba formal llamada *Examen breve del estatus cognitivo*, la cual es opcional y no contribuye a la puntuación de ningún índice. Sin embargo, si una persona obtiene una puntuación baja en esta subprueba, no sería sensato aplicar las subpruebas regulares.

Características técnicas

El WMS-IV fue estandarizado junto con el WAIS-IV. Éste fue un excelente programa de estandarización, como describimos antes. Los índices del WMS usan el ahora conocido sistema de puntuaciones estándar con $M = 100$ y $DE = 15$. Al igual que con el WAIS, la interpretación del WMS comienza con las conclusiones acerca del nivel general de desempeño en comparación con las normas nacionales; luego se continúa con el análisis del perfil y de las discrepancias, es decir, la identificación de fortalezas y debilidades.

El cuadro 8-13 resume la confiabilidad de consistencia interna y la de test-retest de las puntuaciones del WMS-IV. Los datos de consistencia interna provienen de los promedios determinados en cada grupo de edad por separado de la batería de Adultos. La confiabilidad de test-retest se basa en 173 casos con un intervalo promedio de tres semanas.

Cuadro 8-13. Coeficientes de confiabilidad de las puntuaciones del WMS-IV

Tipo de puntuación	Consistencia interna		Test-retest	

	Rango	Mediana	Rango	Mediana
Índice	.93-.96	.95	.81-.83	.82
Subprueba	.82-.97	.90	.64-.77	.74
Proceso	.74-.77	.76	.59-.64	.64

El examen de la información de la confiabilidad sugiere las siguientes generalizaciones. Los índices muestran una consistencia interna excelente y una estabilidad moderada. La mayoría de las subpruebas tiene una consistencia interna adecuada y una estabilidad más bien débil. Las puntuaciones de proceso tienen una consistencia interna ligeramente aceptable y niveles de estabilidad inaceptables para un uso corriente. Recordemos de nuestra descripción de la estructura del WMS que su protocolo demanda el cálculo de muchas diferencias: entre índices, entre puntuaciones del WMS y el WAIS, y así sucesivamente. Aunque los manuales del WMS discuten distintos métodos de examinar estas diferencias y presentan fórmulas para determinar los errores estándar en ellas, los manuales nunca presentan directamente la confiabilidad de estas diferencias. Todos los métodos, en el fondo, exigen interpretar la diferencia entre dos fuentes de información parcialmente no confiables. Como señalamos en el capítulo 4, tales diferencias o contrastes capturan la falta de confiabilidad en las dos puntuaciones implicadas en la comparación y, a menudo, caen debajo de los niveles aceptados de confiabilidad. Último comentario: hay que ser muy, muy cauto al interpretar cualquier tipo de diferencia o contraste entre puntuaciones.

La información sobre la validez del WMS consiste en correlaciones con otras pruebas, diferencias grupales en las puntuaciones promedio y resultados del análisis factorial. En general, esta información apoya el uso del WMS como medida de memoria a corto plazo. Algunos de los numerosos estudios con subgrupos clínicos sugieren la utilidad para propósitos de diagnóstico, es decir, para identificar grupos en los que se espera una pérdida de memoria. ¿Los resultados de análisis factoriales apoyan informar las puntuaciones de los índices (y no la puntuación total de memoria)? Los resultados son algo oscuros; el examen de los datos que se presentan en el manual tanto de los análisis factoriales como de la simple inspección de las correlaciones entre los índices nos permite estar a favor de cualquier cosa, desde una sola puntuación general hasta el conjunto entero de índices.

Discapacidad intelectual y retraso mental: terminología cambiante

Las pruebas de inteligencia de aplicación individual tienen un papel crucial en la definición de individuos con capacidad intelectual muy por debajo del promedio. Como señalamos en nuestra discusión sobre la historia de las pruebas (12-20a»), varios de los desarrollos tempranos en este campo se relacionaron con la identificación de estos individuos. A principios del siglo XX, la definición de "retraso mental" dependía casi de manera exclusiva de las pruebas de inteligencia. Los rangos de CI definían los niveles de retraso; de acuerdo con un sistema muy usado, había tres niveles de retraso, que se definían por los siguientes rangos: 50-70, subnormal; 20-50, imbécil; menos de 20, idiota. En el vocabulario actual, estos términos parecen despectivos y faltos de sensibilidad; sin embargo, en el campo del retraso mental, los términos que ahora encontramos reprobables pueblan, incluso, la historia relativamente reciente. Términos como bobo, defectuoso, lunático, estúpido, subnormal, imbécil e idiota eran comunes en la literatura científica (y legal); eran simples términos técnicos de aquellos días. No debemos suponer que los autores de ese tiempo no tenían corazón; de hecho, muchos de ellos dedicaron toda su vida profesional a mejorar el destino de quienes tenían retraso mental. Sin duda, una futura generación encontrará muy ofensivos algunos de los términos que usamos hoy en el campo de la salud mental. En Scheerenberger (1987), Smith (1997), Editorial Board (1996) y Goode (2002) se pueden encontrar resúmenes útiles de la historia de la definición del retraso mental.

En la mayor parte del siglo XX, el término "retraso mental" se usó para caracterizar individuos con capacidad intelectual muy por debajo del promedio. A principios del siglo XXI, los profesionales se han preocupado cada vez más por la estigmatización de los términos **retraso mental** y **retrasado mental,** por lo que buscaron un término alternativo. La organización profesional de EUA, preocupada principalmente por la definición, clasificación y la atención de estos individuos, eligió el término **discapacidad intelectual**. Por lo tanto, la organización cambió su nombre de *American Association on Mental Retardation* (AAMR [Asociación Americana de Retraso Mental]) a *American Association on Intellectual and Developmental Disabilities* (**AAIDD** [Asociación Americana de Discapacidad Intelectual y del Desarrollo]). Intentó cambiar "retraso mental" por "discapacidad intelectual" en todas sus publicaciones e influir en otras organizaciones para que hicieran lo mismo; sin embargo, el término "retraso mental" está profundamente incrustado en la literatura profesional y en los documentos legales, y parece que no desaparecerá pronto. En esta sección, usamos el término "discapacidad intelectual", excepto cuando citamos documentos que usan de manera explícita el término *retraso mental*. El DSM-5 ha adoptado el término más reciente de discapacidad intelectual en lugar de *retraso mental*, que todavía apareció en el DSM-4 (*American Psychiatric Association,* 2000, 2013).

El concepto de conducta adaptativa

Las definiciones de discapacidad intelectual dependen en gran parte del concepto de **conducta adaptativa**. Primero, exploremos esta noción de manera general y, en la siguiente sección, veremos cómo se incorpora en las definiciones formales de discapacidad intelectual. La conducta adaptativa se refiere a qué tan bien una persona enfrenta la vida cotidiana; otros términos equivalentes que caracterizan esta noción son habilidades adaptativas, funcionamiento adaptativo, habilidades funcionales, funcionamiento cotidiano e, incluso, inteligencia práctica. En fuentes anteriores, **madurez social** y **competencia social** fueron términos comunes. La idea clave es: ¿qué se necesita, al menos en el nivel más sencillo, para arreglárselas en la vida cotidiana? En el cuadro 8-14 se enumeran algunas conductas que definen el funcionamiento adaptativo en distintos niveles.

Cuadro 8-14. Ejemplos de conductas adaptativas en tres niveles distintos

Nivel 1 Alimentarse, vestirse, subir escaleras, decir "hola" y "adiós"
Nivel 2 Decir la hora, hacer cambios, leer palabras sencillas
Nivel 3 Tomar el autobús o el metro, ver las noticias en la TV, comprar la propia ropa

Por supuesto, hay diferencias de edad en las conductas adaptativas; por ejemplo, esperamos que un niño de 10 años de edad sea capaz de amarrarse las agujetas, pero no un niño de 3 años. Esperamos que un joven de 16 años pueda ir a comprar pan, pero no un niño de 5 años. También hay diferencias culturales. Ser capaz de marcar el 911 en una emergencia podría ser importante en una cultura, pero irrelevante en otra. Más adelante, examinaremos pruebas específicas que intentan estimar estas conductas adaptativas.

¡Inténtalo!

Piensa en un niño de 4 años de edad, llamado Frank. Enumera algunas cosas que pienses que Frank necesita ser capaz de hacer para enfrentar su día.

Definición de discapacidad intelectual

Ahora pasemos a las definiciones formales de discapacidad intelectual. En la actualidad, la fuente más común para buscar una definición es la **American Association on Intellectual Disability (AAIDD)**. El libro de AAIDD (2010) *Intellectual Disability: Definition, Classification, and Systems of Support* [Discapacidad intelectual: definición, clasificación y sistemas de apoyo] (a veces referido como el manual de AAIDD, aunque en el título no aparece la palabra *manual*) emplea la siguiente definición (p. 1):

La discapacidad intelectual se caracteriza por limitaciones significativas en el funcionamiento intelectual y en la conducta adaptativa, que se expresa en habilidades adaptativas conceptuales, sociales y prácticas. Esta discapacidad se origina antes de los 18 años de edad.

¡Inténtalo!

Para ver resúmenes de la definición de discapacidad intelectual de AAIDD, así como la variedad de intereses de esa organización, visita el sitio: www.aaidd.org.

Resumen de puntos clave 8-2

Tres criterios de la discapacidad intelectual
(Se deben cumplir los tres)

1. Funcionamiento intelectual significativamente por debajo del promedio
2. Limitaciones en la conducta adaptativa
3. Inicio antes de los 18 años de edad

Podemos notar que hay tres criterios y que los tres se deben cumplir. El *primer criterio* es un funcionamiento intelectual significativamente por debajo del promedio. Aquí hay dos temas: 1) ¿Qué es el funcionamiento intelectual? 2) ¿Qué quiere decir significativamente por debajo del promedio? En la práctica, el funcionamiento intelectual casi siempre se define mediante el desempeño en una prueba individual de inteligencia de las más usadas, como el WISC. La definición de "significativamente por debajo del promedio" ha variado con el paso de los años, pero suele referirse a unidades de desviación estándar en un sistema donde $M = 100$ y $DE = 15$. Lo más común es usar dos *DE* debajo de la media, es decir 70, como el punto de corte. Además, -3 *DE* equivale a 55; -4 *DE*, a 40; y -5 *DE*, a 25; éstos también son puntos de corte que nos ayudan a definir distintos niveles de discapacidad. A menudo se considera un error estándar de medición estimado de 5 puntos del CI en estos puntos de corte, de modo que resultan rangos de 70-75, 50-55, y así sucesivamente, como parte de las definiciones. Esta breve descripción ilustra la importancia de conocer los conceptos de puntuación estándar, unidades de *DE* y error estándar de medición, para comprender el origen de estas definiciones.

La literatura sobre discapacidad intelectual (y retraso mental) se refiere de manera descarada al "CI" y a las "pruebas de CI". Tanto la AAIDD como el DSM-IV (véase más adelante) usan puntuaciones de CI específicas como criterio del funcionamiento intelectual. Así, mientras el resto de la psicología trata de quitar de manera tímida el término *CI* o, al menos, renombrarlo, este campo no manifiesta estos escrúpulos.

El *segundo criterio* de la AAIDD se refiere a las limitaciones en las habilidades adaptativas de distintas áreas, las cuales pueden ser conceptuales, sociales y prácticas. La definición operacional de la discapacidad intelectual exige un desempeño de al menos 2

DE por debajo de la media en, por lo menos, una de las tres áreas o en una puntuación general basada en las tres áreas. (La definición de la AAMR de 1992 incluía 10 áreas, pero la incesante aplicación de los estudios analítico-factoriales, así como en muchas otras aplicaciones, llevó a la reducción a tres áreas.)

El **tercer criterio** es la edad: el padecimiento debe manifestarse antes de los 18 años. Desde el punto de vista práctico, éste no es un tema, por lo común, importante, porque la evaluación suele hacerse antes de esa edad. Sin embargo, desde un punto de vista técnico, es importante señalar que si los dos primeros criterios se cumplen pero el padecimiento no apareció sino hasta, digamos, la edad de 30 años, no se clasificaría como discapacidad intelectual, sino que el diagnóstico sería algo más. Por definición, este padecimiento surge durante los años del desarrollo, definidos operacionalmente como antes de los 18. (Algunas fuentes extienden la definición operacional hasta los 22 años; sin embargo, el punto esencial sigue siendo la existencia de un criterio de desarrollo.)

Una característica importante de la definición de discapacidad intelectual de la AAIDD es la especificación de niveles. Como señalamos antes, los niveles tradicionales dependen principalmente de los rangos de CI; sin embargo, la AAIDD define los niveles en términos de "patrones e intensidades del apoyo necesitado". La AAIDD usa cuatro niveles de apoyo necesitado: intermitente, limitado, extenso y generalizado. El énfasis está en el funcionamiento adaptativo. Los términos se explican por sí mismos y, por supuesto, representan distintos grados de un mismo continuo; por ejemplo, la categoría intermitente significa que la persona necesita ayuda sólo a veces y cuenta con conductas adaptativas básicas. La categoría generalizada significa que la persona depende por completo de alguien más, incluso para funciones tan básicas como ir al baño.

El *Manual diagnóstico y estadístico de los trastornos mentales*, Cuarta edición, Texto revisado (DSM-IV) de la *American Psychiatric Association* (APA, 2000), adopta la definición de tres criterios del retraso mental; sin embargo, usa un sistema diferente para especificar los niveles de gravedad. Podemos notar que estos dependen en gran parte de la puntuación de CI. En la mayoría de los padecimientos, el DSM-IV usa sólo los primeros tres niveles; sin embargo, la categoría "profundo" está muy arraigada en la literatura del retraso mental, así que el sistema del DSM-IV la incorpora. También hay una categoría de "gravedad no especificada", que se usa cuando el juicio clínico sugiere un retraso, pero no ha habido una evaluación formal. El DSM-5 sigue usando los tres criterios y los niveles leve, moderado, grave y profundo, pero pone mayor énfasis en el funcionamiento adaptativo para determinar el nivel de gravedad.

Escalas de la Conducta Adaptativa de Vineland

La medida más usada de la conducta adaptativa son las **Vineland Adaptive Behavior Scales** [Escalas de Conducta Adaptativa de Vineland] (**VABS**) (Sparrow, Balla, & Cicchetti, 1984; Sparrow, Cichetti, & Balla, 2005). Éstas son revisiones de la venerable *Vineland Social Maturity Scale* [Escala de Madurez Social de Vineland] (VSMS; Doll, 1935, 1965). Su autor, Edgar Doll, mediante la elaboración del VSMS y su otro trabajo

en *Vineland* (NJ) *Training School*, ayudó a introducir el concepto de conducta adaptativa. Tanto la escala original como sus más recientes ediciones se suelen conocer simplemente como "el **Vineland**". El Vineland está entre las 20 pruebas más usadas por psicólogos clínicos y tiene el rango más alto en la evaluación funcional/adaptativa y de desarrollo entre psicólogos clínicos y neuropsicólogos (Camara, Nathan, & Puente, 1998, 2000). Lo más importante que se debe aprender aquí es la manera en que el concepto de conducta adaptativa, decisivo en la definición de discapacidad intelectual, puede definirse operacionalmente.

El Vineland usa dos métodos que lo distinguen de las medidas de inteligencia que revisamos antes en este capítulo, por ejemplo, WAIS, WISC y SB5. (Estas diferencias, por lo general, también se aplican a otras medidas de conducta adaptativa.) Primero, el Vineland busca medir el *desempeño típico* más que el máximo; por ejemplo, en el WISC queremos poner a prueba los límites del vocabulario del niño, mientras que en el Vineland, queremos saber qué clases de palabras usa el niño comúnmente. Segundo, el Vineland obtiene información de un **observador externo** (p. ej., un padre) más que por medio de preguntas directas al individuo.

Versiones

El Vineland viene en cuatro versiones; cada una produce puntuaciones de varios dominios y subdominios, así como una puntuación compuesta de Conducta adaptativa. En el cuadro 8-15 aparece un esquema de la estructura del Vineland. ¿Se ve como un modelo jerárquico similar a los que vimos en varias pruebas de capacidad mental que describimos antes en este capítulo? Así es, sólo que *g* no es la cima de la jerarquía.

«227a

Cuadro 8-15. Esquema del Vineland II: estructura jerárquica

Total	Dominio	Subdominio	Ejemplos
Compuesto de conductas adaptativas	Comunicación	Receptivo	
		Expresivo	Usa la comunicación oral
		Escrito	
	Habilidades de la vida cotidiana	Personal	
		Doméstico	Realiza labores de la casa
		Comunidad	
	Socialización	Interpersonal	
		Juego/tiempo libre	Juega
		Habilidades de afrontamiento	
	Habilidades motrices	Gruesas	
		Finas	Usa los dedos para mover objetos
Conducta desadaptada (opcional)	Índice general		
	Reactivos críticos		

El dominio de las Habilidades motrices puede omitirse en el caso de algunos adultos cuando no parece útil. Si se omite, el manual proporciona instrucciones para sustituir puntuaciones al determinar el compuesto de Conducta adaptativa. El dominio de Conducta desadaptada no entra en esta puntuación compuesta.

La aplicación de la **Forma de entrevista de investigación** requiere cerca de 45 min y entre 15 y 30 min más para calificarla. Ésta es la versión estándar, la más usada. El grupo meta de edad va desde el nacimiento hasta los 90 años, una considerable expansión en algunos años a partir de la edición previa. Hay 433 reactivos, pero no todos se usan en todas las entrevistas debido a las reglas basales y de tope. Un examinador entrenado entrevista a un cuidador, por lo general, el padre o la madre del niño, que sea muy cercana a la persona que se está evaluando. De muchas maneras, los procedimientos de aplicación son muy similares a los de las pruebas de inteligencia de aplicación individual, así como a la entrevista semiestructurada que se describe en el capítulo 13. El examinador debe establecer *rapport*, tener íntimo conocimiento de las instrucciones estandarizadas de aplicación, calificar los reactivos de inmediato, explorar cuando se necesite aclarar algo, determinar los niveles basal y de tope dentro de los subdominios, etc.

Cada reactivo identifica una conducta específica. Los reactivos están ordenados en jerarquías dentro de grupos; éstos, a su vez, aparecen dentro de los subdominios. El cuadro 8-16 muestra un reactivo como los del Vineland; éste no es un reactivo real, pero ilustra la estructura y la aplicación de un reactivo.

Cuadro 8-16. Ejemplos ficticios de reactivos relacionados con la conducta adaptativa

Entrevistador: Platíqueme cómo sigue Jack las noticias.
Cuidador: Da una descripción narrativa.
El entrevistador califica cada uno de los siguientes reactivos: 2, 1, 0, N, NS.
El entrevistador explora las descripciones del cuidador, según se requiera, para poder calificar cada reactivo.
Se da cuenta de los eventos actuales importantes, p. ej., una elección presidencial _____
Escucha los informes de noticias en radio o televisión _____
Lee a diario el periódico _____

Con base en la descripción del cuidador, el entrevistador califica un reactivo de acuerdo con este sistema:

2 = Usualmente [hace esto]
1 = A veces o en parte
0 = Nunca [hace esto]
N = No hubo oportunidad [de observar]
NS = No sabe

La **Forma de entrevista** *extensa* incluye todos los reactivos de la Forma de

investigación y reactivos adicionales para proporcionar una descripción más detallada del individuo. Tiene aproximadamente el doble de reactivos y su aplicación tarda el doble de tiempo. Los procedimientos de aplicación son los mismos que los de la Forma de investigación. La Forma de entrevista extensa ofrece información adicional para planear un programa de desarrollo para el individuo y una evaluación de seguimiento del programa.

La **Forma de valoración del padre/cuidador** emplea los mismos reactivos que la Forma de investigación, pero la responde directamente el padre o cuidador. Sin embargo, una persona entrenada en la aplicación del Vineland debe revisar las respuestas con el padre o cuidador después de que ha terminado de responder para llegar a una puntuación final. Esta forma es nueva y apareció en la edición de 2005 del Vineland. El tiempo dirá si es una aportación viable y útil a esta serie de pruebas.

La **Forma de valoración del profesor** (antes Edición para el salón de clases) del Vineland se aplica al profesor, quien la responde con base en sus observaciones en el escenario educativo. No se necesita un entrevistador, y requiere unos 20 min para responderla. El grupo meta es de 3 a 22 años de edad.

En efecto, las cuatro diferentes versiones son cuatro pruebas independientes, cada una con sus propios materiales, procedimientos de elaboración, normas y puntuaciones. Tienen en común la concepción de la conducta adaptativa y, en su mayor parte, los mismos dominios y subdominios. Un instrumento afín son las *Vineland Social-Emotional Early Childhood Scales* [Escalas de la Infancia Temprana Socio-Emocional de Vineland](Vineland SEEC), que está dirigido a niños de 0 a 5 años. Como lo sugiere su nombre, este instrumento se concentra sólo en las áreas del funcionamiento social y emocional.

El Vineland produce la cantidad usual de puntuaciones estandarizadas de los cuatro dominios y el compuesto de Conducta adaptativa: puntuaciones estándar (M = 100, DE = 15), rangos percentiles y estaninas. Podemos notar que el sistema de puntuaciones estándar es el mismo que el de las escalas Wechsler. Los subdominios producen puntuaciones estándar (llamadas puntuaciones de escala v) con M = 15 y DE = 3. De manera extraña, el Vineland proporciona equivalentes de edad de los subdominios, pero no de los dominios, así como rangos percentiles de los dominios, pero no de los subdominios.

Las normas del Vineland tienen dos características especiales. Primero, el Vineland informa los niveles de la conducta adaptativa; a primera vista, éstos pueden parecer que son las definiciones de criterio del funcionamiento adaptativo, pero no lo son. Son estrictamente con referencia a una norma, definidos en unidades DE o rangos percentiles (RP) de la siguiente manera:

Nivel adaptativo	**Unidades DE**	**Unidades RP**
Bajo	−2 o menos	2 o menos
Moderadamente bajo	de −1 a −2	3-17
Adecuado	de +1 a −1	18-83

Moderadamente alto	de +1 a +2	84-97
Alto	+2 o más	98 o más

El término adecuado para la categoría intermedia suena a la interpretación con referencia a un criterio. Se trata de un uso desafortunado, ya que no parece haber bases para declarar esta categoría, ni ninguna otra, como adecuada o inadecuada. También debemos notar que la categoría intermedia (adecuada), que cubre del percentil 18 al 83, es muy amplia.

La segunda característica especial de las normas del Vineland se relaciona con el dominio de la Conducta desadaptada. En realidad, toda esta parte de la prueba merece un comentario especial. Este dominio no forma parte del compuesto de la Conducta adaptativa; además, es opcional. Las puntuaciones altas en el dominio de la Conducta desadaptada son indeseables, mientras que el resto de las áreas son deseables. El dominio de Conducta desadaptada incluye 11 reactivos "internalizantes", 10 "externalizantes", 15 de "otros" y 14 "críticos". Se proporcionan puntuaciones separadas de escala *v* para Internalizante, Externalizante y el índice (total) de Conducta desadaptada, que incluye los reactivos internalizante, externalizante y otros. Los reactivos críticos permanecen aislados, pues no entran en ninguna puntuación. Este dominio proporciona los niveles de desadaptación, definidos de la siguiente manera:

Nivel de desadaptación	**Unidades DE**	**Puntuación de escala *v***
Promedio	Hasta +1	Menos de 18
Elevado	De +1 a +2	18-20
Clínicamente significativo	Más de +2	21-24

La edición de 1984 del Vineland proporcionaba normas abreviadas independientes de varios grupos clínicos (p. ej., niños con perturbación emocional, adultos con retraso mental ambulatorios). La edición de 2005 no proporciona tales normas.

Características *técnicas*

Las características técnicas del Vineland son, en conjunto, excelentes. Las normas se basan en lo que parecen ser muestras bien definidas y elegidas de manera competente. El manual informa la confiabilidad de consistencia interna, de test-reteste interjueces. En general, la confiabilidad de consistencia interna y la de test-retest son excelentes, aunque con algunas excepciones. Gran parte de la confiabilidad interjueces está entre .70 y .80, a veces un poco abajo, lo que nos advierte que se debe tener cuidado al interpretar las puntuaciones. El Vineland tiene un información excelente sobre la validez, incluyendo correlaciones con otras pruebas, análisis factoriales y desempeño de subgrupos pertinentes. Los análisis factoriales brindan apoyo a la estructura de dominios del

Vineland, aunque se debe admitir que una solución de un factor (conducta adaptativa general) proporciona un ajuste casi tan bueno a los datos. Como suele suceder, la interpretación de todos los datos del manual de los numerosos contrastes grupales y correlaciones con una multitud de otras pruebas presenta un desafío abrumador para el lector.

Otras escalas de adaptación

El Vineland es el instrumento por excelencia para medir el funcionamiento adaptativo; sus primeras ediciones, casi sin más ayuda, definieron el campo entero de la conducta adaptativa y, en el proceso, moldearon de manera significativa las definiciones contemporáneas de discapacidad intelectual. Sin embargo, existen numerosas medidas de la conducta adaptativa que, en general, tienden a seguir los patrones básicos establecidos por el Vinelad: se concentran en las habilidades y conductas cotidianas, evalúan el desempeño típico y se basan en informes de otras personas. Las alternativas difieren del Vineland en cuestiones como los dominios específicos evaluados y el nivel de detalle. Estos instrumentos también muestran una gran variación en la riqueza de la investigación que los respalda.

¡Inténtalo!

Para examinar distintas medidas de conducta adaptativa, visita el sitio de ETS Test Collection (http://www.ets.org/test_link/find_tests/). Introduce como palabras clave adaptive behavior (conducta adaptativa), adaptive functioning (funcionamiento adaptativo) o social maturity (madurez social).

Pruebas para la infancia temprana

No revisaremos un ejemplo específico de una prueba diseñada para niños muy pequeños; sin embargo, discutiremos brevemente algunas características importantes de este tipo de pruebas. Hay tres puntos importantes que considerar. Primero, las categorías generales de los reactivos para estas edades son similares a las de los reactivos que se usan para edades mayores: palabras, memoria, tareas psicomotrices, material cuantitativo, entre otras. Sin embargo, las tareas son de un nivel tan sencillo que no es claro si miden las mismas dimensiones que en edades mayores; por ejemplo, ser capaz de reconocer la diferencia entre canicas y sólo una canica puede no corresponder a la misma dimensión que resolver un problema aritmético con palabras, aunque ambos reactivos sean de naturaleza cuantitativa. Reconocer el significado de "mano" puede no corresponder a la misma dimensión que definir "arrogante", aunque ambos tengan que ver con el significado de las palabras.

Segundo, el énfasis en estas jóvenes edades recae en el estatus del desarrollo más que en la inteligencia. De hecho, tal vez no tenemos una idea muy clara de qué significa la inteligencia en un niño, digamos, de 2 años de edad. Quizá estos dos puntos ayuden a explicar el insignificante poder predictivo de estas pruebas. Tercero, las pruebas para edades muy tempranas sirven primordialmente como exploración. Para la población general, no hay una correlación alta entre la inteligencia medida en edades posteriores (p. ej., de 6 años en adelante) y las medidas en edades muy tempranas; sin embargo, hay una alta correlación de casos en la parte más baja de la distribución. Por ejemplo, el retraso moderado y grave se manifiestan en edades tempranas; de ahí el interés en explorar casos que pueden requerir una evaluación más detallada. En el caso de niños cuyo desarrollo sigue el curso normal, no vale la pena tratar la inteligencia en edades tempranas. Entre las medidas más populares del desarrollo temprano están las *Escalas Bayley de Desarrollo Infantil* para edades de 1 a 42 meses, el *McCarthy Scales of Children's Abilities* [Escalas McCarthy de Capacidades de Niños] para edades de 2-6 a 8-6 y el clásico *Gessell Developmental Schedules* [Escalas de Desarrollo de Gessell] para edades de 4 semanas a 6 años.

Otras aplicaciones

En las secciones anteriores, observamos cómo un aspecto del funcionamiento intelectual general (memoria) puede explorarse con mayor detalle. En otra sección, vimos cómo se puede emparejar una medida de funcionamiento intelectual general con la medición de otro constructo (conducta adaptativa) para ayudar a definir un padecimiento (discapacidad intelectual). Estos ejemplos podrían ampliarse casi indefinidamente; se podrían incluir las medidas aplicables a los problemas de aprendizaje, déficit de atención, demencia de tipo Alzheimer, creatividad, deterioro auditivo, genialidad cuantitativa. Está de más decir que el espacio no nos permite tratar todas estas aplicaciones en un texto introductorio; sin embargo, podemos señalar que, sin importar la aplicación, las preguntas siempre son las mismas.

- ¿Cómo conceptualizamos el tema (p. ej., memoria, dislexia)?
- ¿Qué pruebas o combinación de pruebas (e informa-ción de otro tipo de fuentes) podrían ser útiles?
- ¿La prueba proporciona información confiable?
- ¿Qué evidencia existe sobre la validez de la prueba?
- ¿Las normas son reflejo de un grupo bien definido?

Si seguimos esta línea de pensamiento, seremos capaces de acercarnos a cualquier área de interés con cierto éxito. En el capítulo 10, examinaremos aplicaciones adicionales a otros padecimientos.

¡Inténtalo!

Para ver cómo podrían aplicarse las pruebas de capacidad mental a una de las áreas mencionadas antes, introduce uno de los términos (p. ej., dislexia) como palabra clave en PsychINFO. Enumera las pruebas que se usen para ayudar a evaluar esa área.

Tendencias en las pruebas individuales de inteligencia

Se pueden detectar varias tendencias en la naturaleza y uso de las pruebas de capacidad mental de aplicación individual en los últimos años. Las ediciones más recientes de las pruebas más usadas, así como el surgimiento de nuevas pruebas, muestran estas tendencias, algunas de las cuales también caracterizan a las pruebas de capacidad mental de aplicación grupal, pero aquí nos concentraremos en las de aplicación individual. Identificaremos seis tendencias, algunas de las cuales tienen, en parte, distintos elementos entre ellas.

Primero, las pruebas usan cada vez más alguna versión del *modelo jerárquico* de la inteligencia como su marco teórico; estos modelos no se usan de manera rígida, sino que son una guía para la elaboración e interpretación de las pruebas. Ahora son comunes las referencias a los modelos jerárquicos de Vernon, Cattell y Carroll en los manuales de las pruebas de capacidad mental de aplicación individual. De hecho, el usuario de la prueba debe estar familiarizado con estos modelos para seguir las discusiones de la interpretación de la prueba en el manual.

Segundo, entre las pruebas integrales, hay una tendencia hacia la *mayor complejidad* tanto en su estructura como en la manera en que se usan las puntuaciones, lo cual, en parte, se debe al uso de los modelos jerárquicos, pues inevitablemente llevan a producir más puntuaciones. Las demandas legales recientes también llevan a la necesidad de más puntuaciones; revisaremos algunas de estas demandas en el capítulo 16. Sin embargo, aquí podemos señalar que si la definición de retraso mental se refiere al déficit en al menos dos de 10 áreas, sin duda, esto sugiere que uno debe evaluar las 10 áreas. Además, identificar los problemas de aprendizaje depende en gran parte de comparar las puntuaciones de distintas áreas, lo que sugiere el uso de un instrumento de puntuaciones múltiples, de los cuales el Stanford-Binet ofrece, quizá, el ejemplo más claro. Mientras que esta prueba ofrece sólo una puntuación global en las ediciones que se publicaron durante 70 años (1916-1986), la de 1986 produce una puntuación total, cuatro subpruebas importantes y 15 subpruebas secundarias. Las ediciones más recientes de las escalas Wechsler también han agregado puntuaciones. Sin embargo, más allá de la producción de más puntuaciones, las pruebas se han vuelto más complejas en la manera en que se usan las puntuaciones. Los manuales de las pruebas contienen recomendaciones de más comparaciones entre las puntuaciones, y los sistemas de calificación por computadora facilitan la multiplicación de comparaciones. La interacción de todos eque va a interpretar los resultados deintérprete de las pruebas. Recordemos, por ejemplo, que el error estándar de la diferencia entre dos pruebas no es sólo la suma de los errores estándar de cada una de ellas. Éste es sólo un ejemplo del hecho de que la proliferación de puntuaciones requiere una interpretación más sofisticada. La disponibilidad de informes narrativos generados por computadora, que con facilidad

puede prolongar las comparaciones, puede ser de gran ayuda para el usuario, pero demanda cautela extra.

Tercero, los materiales para la enseñanza correctiva acompañan cada vez más a las pruebas. Es decir, una vez que se ha determinado un perfil de fortalezas y debilidades, hay materiales que buscan aprovechar las fortalezas y corregir las debilidades. Esto es un derivado directo de las pruebas de puntuaciones múltiples aplicadas a personas con problemas de aprendizaje, TDAH, discapacidad intelectual y otros padecimientos de ese tipo. Aunque esta práctica tiene una larga historia en las pruebas de aprovechamiento, es un desarrollo notable para las pruebas de capacidad mental. No hemos examinado estos materiales para la enseñanza, porque nos llevaría demasiado lejos; sin embargo, es una tendencia inconfundible.

Cuarto, aunque las escalas tradicionales como el WISC y el WAIS aún predominan entre las pruebas de capacidad mental de aplicación individual, parece estar en aumento el uso de instrumentos más breves debido principalmente a la mayor demanda de eficiencia en los servicios de la industria de la salud (véase Daw, 2001; Piotrowski, 1999; Thompson *et al*., 2004). Si un tercero paga el servicio, no querrá pagar por una prueba de 90 min si una de 10 puede proporcionar la información necesaria. Este tipo de demandas se ha extendido ahora por la mayor parte de la industria de la salud. La provisión de servicios psicológicos es el simple reflejo de una tendencia más amplia; una razón secundaria de ésta puede ser la mayor complejidad ya mencionada, es decir, los instrumentos más extensos pueden volverse demasiado complejos.

Quinto, casi sin excepción, las pruebas más usadas de esta categoría cuentan con normas excelentes. El proceso de elaboración de normas nacionales para las pruebas se ha vuelto muy sofisticado y estandarizado; aunque algunas pruebas menos conocidas pueden depender del muestreo por conveniencia, esta práctica no es típica de las pruebas más usadas. No sólo el proceso de estandarización, sino también la descripción de este proceso en los manuales han alcanzado un alto nivel de excelencia; en las primeras ediciones de estas pruebas, sólo se dedicaban una o dos páginas a esta información, pero ahora ocupa comúnmente docenas de páginas.

Sexto, la atención al sesgo de las pruebas se ha vuelto explícita en su preparación. Los reactivos de las últimas ediciones de estas pruebas de manera rutinaria están sujetos a revisión por parte de paneles de minorías representativas. Las editoriales de pruebas, por lo común, emplean procedimientos estadísticos para detectar sesgos en los reactivos, como lo describimos en las páginas 163-165a». Además, la discusión de la interpretación de las pruebas pone mucha mayor atención que en épocas pasadas a los posibles efectos de factores ambientales y culturales. Prácticamente nadie afirmaría hoy que el desempeño en estas pruebas sólo es atribuible a la capacidad "innata". Ha aumentado enormemente la preocupación por los grupos minoritarios y las personas con discapacidades, lo cual ha motivado mucho del desarrollo en esta área. En el capítulo 6 se pueden encontrar más detalles sobre este tema.

Resumen de puntos clave 8-3

Tendencias recientes en las pruebas de inteligencia de aplicación individual

- Uso de un modelo jerárquico en la estructura de la prueba
- Mayor complejidad
- Provisión de materiales correctivos
- Mayor uso de las pruebas breves
- Estandarización más sofisticada
- Atención al sesgo de las pruebas
- Mayor frecuencia de la revisión de pruebas

Por último, señalamos el aumento en la frecuencia de la revisión de pruebas individuales de inteligencia. Regresemos a la década de 1950 y consideremos estos datos. El periodo entre la aparición del WAIS y el WAIS-R fue de 26 años; entre el WISC y el WISC-R, 25 años; entre el PPVT y el PPVT-R, 22 años. Las nuevas ediciones de estas pruebas ahora aparecen más o menos cada 10 años. ¿Por qué esto es importante? Porque se espera que el psicólogo sea competente en el uso de las últimas ediciones de estas pruebas. La ética profesional, tratada en detalle en el capítulo 16, demanda una continua educación; antes, el psicólogo podía progresar a lo largo de toda su carrera conociendo sólo una o dos ediciones de estas pruebas, pero ahora aparecen cuatro o cinco durante su carrera. Esto resalta la importancia de ser capaz de evaluar información nueva acerca de temas como confiabilidad, validez y normas. Así, es necesario estudiar los libros de texto sobre pruebas psicológicas con diligencia.

Resumen

1. Los psicólogos emplean pruebas de inteligencia de aplicación individual en una gran variedad de aplicaciones prácticas.
2. Las pruebas individuales de inteligencia tienen las siguientes características en común: son de aplicación individual, requieren un entrenamiento avanzado para aplicarlas, cubren un amplio rango de edades y capacidades, deben establecer el *rapport*, emplean el formato de respuesta libre, los reactivos se califican de inmediato, su aplicación toma cerca de una hora y ofrecen la oportunidad de observar al examinado.
3. Muchas pruebas usan reactivos de estas categorías: vocabulario, relaciones verbales, información, significado (comprensión), razonamiento aritmético, memoria a corto plazo, patrones de formas y habilidades psicomotrices. Algunas pruebas sólo usan una o pocas de estas categorías.
4. Las escalas Wechsler constituyen una familia de pruebas; varias de ellas están entre las más usadas en psicología.
5. La Escala Wechsler de Inteligencia para Adultos (WAIS), ahora en su cuarta edición, presenta numerosas puntuaciones de subpruebas, cuatro índices y un CI Total. Examinamos la naturaleza de las subpruebas en detalle. Los estudios de estandarización, confiabilidad y validez del WAIS son de alta calidad. En la interpretación de las puntuaciones del WAIS se le da gran importancia al análisis del perfil y de las discrepancias. Estos procedimientos demandan especial cautela.
6. La Escala Wechsler de Inteligencia para Niños (WISC), ahora en su cuarta edición, tiene una estructura, propósito y calidad técnica muy similares a los del WAIS. Sin embargo, existen algunas diferencias, en especial en la lista de subpruebas.
7. La quinta edición del venerable Stanford-Binet hizo cambios significativos en la estructura respecto de sus antecesores. Aún proporciona una puntuación general, pero ahora también produce puntuaciones en cuatro áreas importantes (capacidad verbal, cuantitativa, abstracta/visual y memoria a corto plazo) y en una gran cantidad de áreas específicas.
8. Para algunos propósitos, una medida breve de capacidad mental es suficiente. Un ejemplo excelente de tal medida es la Prueba de Vocabulario en Imágenes Peabody (PPVT), que se basa por completo en el vocabulario auditivo. Su aplicación requiere sólo 15 min y emplea un formato de opción múltiple. Tiene correlaciones altas con medidas de inteligencia más extensas y definidas de manera global.
9. La Escala de Memoria de Wechsler (WMS) ilustra cómo se puede medir en profundidad un solo aspecto de la capacidad mental. Psicólogos clínicos y neuropsicólogos usan mucho el WMS a causa de su sensibilidad para las funciones de la memoria en numerosos padecimientos debilitantes.
10. La discapacidad intelectual (antes llamada retraso mental), alguna vez definida casi exclusivamente con base en el CI, ahora depende, en parte, de la noción de conducta

adaptativa.

11. Las medidas más usadas de conducta adaptativa son las ediciones del "Vineland". Esta prueba intenta medir el desempeño típico por medio de informes de individuos cercanos a la persona que se está evaluando.

12. Identificamos siete tendencias en las pruebas de inteligencia de aplicación individual: 1) uso de modelos jerárquicos de inteligencia para determinar la estructura de las pruebas, 2) aumento en la complejidad de la estructura de la prueba, número de puntuaciones y métodos para informar, 3) provisión de materiales correctivos para dar seguimiento a las puntuaciones bajas, 4) crecimiento en el uso de pruebas breves, principalmente como resultado de la presión por parte de instituciones de salud, 5) estandarización más sofisticada de las pruebas, 6) mayor atención al sesgo de las pruebas durante el proceso de su elaboración y 7) mayor frecuencia de la revisión de estas pruebas.

Palabras clave

AAIDD
CIE
CIT
CIV
conducta adaptativa
discapacidad intelectual
memoria demorada
memoria inmediata
PPVT
puntuación de índice
rapport
reglas de inicio y discontinuación
retención de dígitos
Stanford-Binet
VABS
Vineland
WAIS
WISC
WMS
WPPSI

Ejercicios

1. Observa la lista de subpruebas del WAIS que aparece en el cuadro 8-5. Asigna cada una a uno de los estratos del nivel 2 del modelo jerárquico de Carroll (184b»). Compara tus asignaciones con las de alguien más.
2. Entre los reactivos que se usan por lo común en las pruebas individuales de inteligencia están vocabulario, información y problemas aritméticos expresados con palabras. Para cada una de estas áreas, prepara tres reactivos adecuados para niños de 6 años de edad.
3. La American Association on Intellectual and Developmental Disabilities (AAIDD) es la principal fuente de la definición de discapacidad intelectual. Esta organización también trabaja en pro de iniciativas legislativas. Para revisar los últimos desarrollos en la AAIDD, entra a www.aaidd.org.
4. Con ayuda del sitio de ETS Test Collection (http://www.ets.org/test_link/find_tests/), identifica tres pruebas de conducta adaptativa, las que sea, y llena este cuadro:

Prueba	Nombre de la prueba	Rango de edad	Puntuaciones
1			
2			
3			

5. En las Escalas de Conducta Adaptativa de Vineland (véase 227a»), observa los subdominios de comunicación receptiva y expresiva. Identifica dos ejemplos de cada área de habilidad que puedan usarse en la prueba. Recuerda que estas habilidades deben mostrarse de manera típica y ser observables para la persona a la que se entrevista.
6. Vas a aplicar el WISC a un niño de 6 años de edad. Lo primero que tienes que hacer es establecer el *rapport*. ¿Qué podrías decir para lograrlo? ¿Y si se tratara de un joven de 16 años?
7. Vuelve a observar los reactivos muestra que aparecen en el cuadro 8-2. Por cada categoría, escribe dos reactivos que puedan usarse en una prueba individual de inteligencia: uno para la edad de 6 años y otro para la edad de 20.
8. En la figura 8-8, al principio de la sección sobre la Escala de Memoria de Wechsler, se ilustró la manera en que un área (memoria) dentro del funcionamiento intelectual general puede extenderse en una prueba más detallada. Toma alguna otra área, por ejemplo, capacidad verbal o cuantitativa, y haz una lista de subpruebas que podrías crear para proporcionar una medición más detallada de esta área. Compara tu lista con la de alguien más.
9. Vuelve a observar el cuadro 8-7. En él se muestran las porciones de los cuadros de

normas de Vocabulario y Diseño con cubos del WAIS de dos grupos de edad. Con base en lo que sabes acerca de la distribución normal y las puntuaciones estándar (la puntuación escalar en la columna izquierda del cuadro), dibuja las distribuciones de los dos grupos de edad mostrando su superposición. Haz dos dibujos: uno que muestre los dos grupos en Vocabulario y otro, en Diseño con cubos.

CAPÍTULO 9
Pruebas grupales de capacidad mental

Objetivos

1. Identificar las principales categorías y aplicaciones de las pruebas grupales de capacidad mental más usadas.
2. Enumerar las ocho características en común de las pruebas grupales de capacidad mental, sobre todo en contraste con las de aplicación individual.
3. De cada una de las siguientes pruebas, dar el nombre completo, propósito, grupo meta, escala de las puntuaciones y un breve resumen de su confiabilidad y validez: OLSAT, SAT, ACT, GRE-G, ASVAB y Wonderlic.
4. Describir los intentos de construir pruebas de capacidad mental que sean neutrales en términos culturales y dar ejemplos de los tipos de reactivos que se emplean en dichas pruebas.
5. Discutir las seis generalizaciones respecto de las pruebas grupales de capacidad mental.

Algunos casos

- La maestra Vásquez da clases al quinto grado de la escuela St. Cecilia en Exeter. Este año, ella es nueva en la escuela, por lo que no conoce a los niños. Al final del cuarto grado, los alumnos respondieron una prueba estandarizada de aprovechamiento y otra de capacidad mental. La maestra Vásquez planea hacer trabajar al máximo a todos sus estudiantes; sin embargo, está preocupada por identificar dos clases de alumnos. Primero, ¿tiene alumnos muy brillantes con un nivel mediocre de aprovechamiento? Si es así, ella quiere encontrar una manera especial de motivarlos para que este año, "en verdad, sobresalgan." Segundo, ¿tiene alumnos con una capacidad mental muy baja? Mediante su enfoque en el que no hay lugar para el "no tiene sentido estudiar", quiere estar segura de que estos alumnos no se desanimen. La maestra Vásquez imprime copias de las puntuaciones de las pruebas con estas preguntas en mente.
- El Colegio Ivy se precia de su ya tradicional plan de estudios liberal: muchas mucha lectura, muchas tareas de escritura. También hay muchos aspirantes provenientes de preparatorias privadas y de escuelas públicas suburbanas de alto nivel socioeconómico. Al Colegio Ivy le gustaría atraer también a estudiantes de escuelas rurales y de nivel socioeconómico bajo, que nunca hayan tomado cursos avanzados, viajado a Europa ni visitado un museo. Al mismo tiempo, el Colegio Ivy quiere asegurarse de que estos alumnos tengan la capacidad básica para tener éxito en este plan de estudios dinámico. ¿Hay alguna prueba de "capacidad básica" que ayude a elegir a estos alumnos?
- Este año, el Ejército de EUA reclutará a 100 000 jóvenes, hombres y mujeres, que se desempeñarán en 300 trabajos, que incluyen operador de radar, bombero, analista de inteligencia, mecánico en aviación, especialista en servicios de comida, etc. ¿Hay algo acerca de las capacidades mentales de estos reclutas que ayude a asignarlos a los distintos trabajos? ¿Qué incluirías en una prueba para medir sus capacidades mentales?

Todos estos casos ilustran situaciones que requieren la aplicación de pruebas de capacidad mental de aplicación grupal. En este capítulo, examinaremos las características de estas pruebas y describiremos varios ejemplos específicos.

Usos de las pruebas grupales de capacidad mental

Las pruebas de capacidad mental de aplicación grupal que examinamos en este capítulo están entre las más usadas de todas las pruebas. Aunque no hay disponible ningún conteo, parece seguro estimar que cerca de 50 millones de estas pruebas se aplican cada año sólo en EUA. Para cuando un individuo típico llega a la edad adulta, al menos en EUA y en muchos otros países occidentales, se le han aplicado media docena de estas pruebas o más. Este tipo de pruebas se dividen en cuatro grupos principales; con base en esta división organizamos el esquema de este capítulo.

El primer uso importante se realiza en las escuelas primarias y secundarias, donde las pruebas grupales de capacidad mental se aplican a menudo junto con una prueba estandarizada de aprovechamiento como parte de los programas de evaluación escolares. Se comparan los resultados de las pruebas de capacidad mental y de aprovechamiento para determinar si el aprovechamiento del alumno es razonablemente consistente con su capacidad mental. Las pruebas de capacidad mental también pueden emplearse como uno de varios criterios para elegir alumnos en programas especiales, por ejemplo, en uno para alumnos sobredotados o de educación especial.

El segundo uso importante de las pruebas de capacidad mental de aplicación grupal es para predecir el éxito en la universidad o en programas para graduados o profesionales. Entre estas pruebas se encuentran el SAT, ACT, GRE y LSAT, acrónimos que son conocidos para la mayoría de los lectores de este libro. Esta categoría se divide en dos. Primero, están las dos pruebas predominantes en la selección y ubicación de alumnos universitarios: el SAT y el ACT. En segundo lugar se encuentran numerosas pruebas empleadas para la selección de alumnos para programas de posgrado y de profesionalización.

El *tercer* uso importante de estas pruebas es para la selección de aspirantes a un trabajo o la ubicación en contextos militares o de negocios. Cada recluta del Ejército de EUA responde al menos una prueba de capacidad mental. Recordemos del capítulo 1 (16a») que la evaluación de grandes cantidades de reclutas del Ejército estimuló el desarrollo de la primera prueba de capacidad mental de aplicación grupal. Muchos negocios también usan pruebas de capacidad mental como un criterio para elegir a sus empleados.

Cuarto, las pruebas de capacidad mental de aplicación grupal se utilizan mucho para la investigación en ciencias sociales y de la conducta. Algunas de estas investigaciones se relacionan directamente con la naturaleza de la capacidad mental y su relación con otras variables, como edad, ingresos, resultados educativos o variables de personalidad. En otros casos, se incluye una medida de inteligencia simplemente para describir la muestra que se emplea en un proyecto de investigación.

Resumen de puntos clave 9-1

Principales usos de las pruebas grupales de capacidad mental

- En escuelas primarias y secundarias, junto con pruebas de aprovechamiento
- Para predecir el éxito en:
- la universidad
- programas de posgrado y de profesionalización
- Elección de trabajo o ubicación en el Ejército y los negocios
- Investigación

Características en común de las pruebas grupales de capacidad mental

Las pruebas de capacidad mental de aplicación grupal comparten varias características. De particular importancia son sus diferencias en relación con las pruebas de aplicación individual que describimos en el capítulo anterior.

Primero, lo más evidente es que estas pruebas se pueden aplicar a grupos grandes. En teoría, no hay límites para el tamaño del grupo, pero en una situación típica, hay entre 20 y 50 examinados por cada aplicador. Se pueden aplicar a grupos más grandes, de varios cientos, con un aplicador principal y varios censores. El aplicador principal lee las instrucciones mientras los censores circulan para mantener el orden, evitar las trampas, etc. Podemos notar que cualquier prueba que se pueda aplicar a un grupo grande también se puede aplicar a un individuo, lo cual a veces se hace.

¡Inténtalo! «234a

Haz una lista de las pruebas de capacidad mental de aplicación grupal que has respondido en los últimos cinco años. Piensa si estas pruebas coinciden con las características que bosquejamos en esta sección.

Segundo, las pruebas de aplicación grupal casi siempre están integradas por reactivos de opción múltiple, por lo que son susceptibles de ser calificadas de manera automatizada. Estas características en particular (reactivos de opción múltiple y calificación automatizada) distinguen a las pruebas de aplicación grupal de las de aplicación individual. Desde luego, hay excepciones, aunque son muy pocas.

Tercero, a pesar de las diferencias en el formato, el contenido de las pruebas individuales y grupales es muy similar. En las pruebas grupales, encontramos vocabulario, analogías verbales, razonamiento aritmético, información y muchas otras clases de reactivos que también encontramos en las pruebas individuales. La semejanza en el contenido no es una sorpresa; recordemos que Otis creó la primera prueba de capacidad mental de aplicación grupal tratando de emular la de Binet. Además, los autores de pruebas individuales y grupales recurren a la misma base de investigaciones acerca de la naturaleza de la inteligencia para crear sus pruebas.

El cuadro 9-1 «235a muestra ejemplos de reactivos con formatos de respuesta libre y de opción múltiple que aparecen a menudo en las pruebas individuales y grupales. Los reactivos de respuesta libre provienen del cuadro 8-2, donde aparecen como ejemplos de reactivos que podrían aparecer en pruebas de inteligencia de aplicación individual. La diferencia más obvia entre estos dos formatos es la presencia de opciones de respuesta en el caso de la versión de opción múltiple, pero no es tan obvio el hecho de que el reactivo se le dice de manera oral en el formato de respuesta libre en la aplicación

individual, mientras que en el formato de opción múltiple en la aplicación grupal es el individuo quien lee el reactivo (excepto en las pruebas para grados inferiores al tercero).

Cuadro 9-1. Ejemplos de reactivos con formato de respuesta libre y de opción múltiple

Usados típicamente en	*Pruebas de aplicación individual*	*Pruebas de aplicación grupal*
Formato	*Respuesta libre*	*Opción múltiple*
Vocabulario	¿Qué significa arrogante?	Arrogante significa a) tacaño b) obeso c) altivo d) malévolo
Relaciones verbales	El padre es al hijo lo que la madre es a...	El padre es al hijo lo que la madre es a a) la hermana b) la sobrina c) la hija d) el tío
Razonamiento aritmético	Jim compró 5 lápices a 12 pesos cada uno y 2 libretas a 80 pesos cada una. ¿Cuánto pagó?	Jim compró 5 lápices a 12 pesos cada uno y 2 libretas a 80 pesos cada una. ¿Cuánto pagó? a) $220 b) $120 c) $232 d) $99

Hay dos excepciones a esta generalización acerca de la semejanza en el contenido de las pruebas grupales e individuales. Primero, las pruebas grupales no suelen incluir reactivos que midan la memoria a corto plazo ni, segundo, reactivos que requieran de manipulación de cubos, rompecabezas u otro tipo de materiales. La aplicación grupal, por lo común, excluye el uso de estos dos tipos de reactivos.

Cuarto, hay un límite de tiempo fijo y un número fijo de reactivos en estas pruebas. Por ejemplo, una prueba grupal puede tener 80 reactivos y un límite de tiempo de 50 min. Todos los examinados ven todos los reactivos, aunque no todos los respondan, y trabajan con el mismo límite de tiempo. Estos arreglos contrastan con los de las pruebas de aplicación individual, con sus reglas de inicio y discontinuación que llevan a tiempos variables de aplicación para cada individuo. Algunas pruebas de aplicación grupal están disponibles en el modo de aplicación adaptadas para computadora, en el cual puede no haber un límite de tiempo fijo o un número fijo de reactivos. Sin embargo, aún sucede que la gran mayoría de pruebas de aplicación grupal emplean el formato convencional. Los tiempos de aplicación de las pruebas grupales muestran una distribución bimodal peculiar, pues en muchas de ellas el tiempo va de 45 a 60 min y en otras de 2 hrs y media a 3 hrs.

En términos del número de puntuaciones, las pruebas grupales son similares a las individuales. Suelen producir una puntuación total y varias subpuntuaciones, por ejemplo, verbales y cuantitativas, o verbales y no verbales. Algunas pruebas proporcionan hasta 10 o 12 subpuntuaciones, cada una de las cuales tiene, por lo común, de 25 a 50 reactivos y requiere entre 20 y 50 min para su aplicación. Desde luego, hay variaciones a estos patrones, pero éstos suelen ser bastante constantes en una gran cantidad de pruebas de capacidad mental de aplicación grupal.

Entre las pruebas grupales de capacidad mental más usadas, la investigación en que se basan las normas, la equiparación, la confiabilidad, etc., es *muy amplia*, por lo general, más que en el caso de las pruebas de aplicación individual. Mientras que el grupo de estandarización para una prueba individual puede implicar 2000 casos, el de las pruebas grupales puede implicar 200 000 o incluso un millón de casos.

Resumen de puntos clave 9-2

Características en común de las pruebas grupales de capacidad mental «236a

- Se pueden aplicar en grupos grandes
- Reactivos de opción múltiple, calificados de manera automatizada
- El contenido, por lo general, es similar al de las pruebas individuales
- Límite de tiempo fijo, número fijo de reactivos
- Tiempo de aplicación: 1 o 3 hrs
- Puntuación total más varias subpuntuaciones
- Amplia base de investigaciones
- Propósito principal: predicción

Prácticamente todas las pruebas de esta categoría tienen la *predicción como su propósito principal*. Algunas buscan predecir el desempeño en la escuela, desde los primeros grados de la escuela primaria hasta en los programas de posgrado o profesionalización. Desde luego, desde un punto de vista práctico, la predicción a menudo se traduce en selección, por ejemplo, para un programa de posgrado. Otras pruebas buscan predecir el éxito en un trabajo; de ahí que los estudios de **validez predictiva** tienen un lugar de honor en la literatura de la investigación sobre estas pruebas. El efecto de la restricción en el rango de los coeficientes de correlación es particularmente agudo en estos estudios de validez predictiva. Con el fin de interpretar los datos de la validez predictiva de estas pruebas, en especial las que se encuentran en el rango superior de capacidad, quizá el lector quiera revisar este tema (restricción en el rango) en las páginas 80-83. El efecto de la confiabilidad imperfecta en el criterio (véase pp. 117-118a») también es importante en estos estudios. Un propósito adicional de muchas de las pruebas es ayudar en la ubicación en escenarios educativos o laborales para maximizar el éxito, pero la predicción o selección sigue siendo el propósito principal.

Pruebas de capacidad mental en programas de evaluación escolar

Como señalamos antes, un uso importante de las pruebas de capacidad mental de aplicación grupal tiene lugar en los programas de evaluación escolar. Aunque hay un gran número de este tipo de pruebas, sólo unas pocas caen en esta categoría, pero son muy usadas. El cuadro 9-2 enumera las principales entradas de esta categoría: la **Prueba de Capacidad Escolar Otis-Lennon (OLSAT)**, el *InView* y el *Cognitive Abilities Test*. Cada una de estas pruebas está apareada con una de las principales baterías estandarizadas de aprovechamiento, como lo veremos más adelante en el capítulo 11. Abajo describimos una de estas series de pruebas, el Otis-Lennon. Las tres pruebas que aparecen en el cuadro 9-2 son muy parecidas en muchos sentidos; de ahí que muchas de nuestras observaciones acerca del Otis-Lennon se puedan aplicar a las otras pruebas del cuadro 9-2.

Cuadro 9-2. Pruebas grupales de capacidad mental usadas en programas de evaluación escolar

Prueba	Niveles/Grados	Puntuaciones	Editorial	Relacionado con
Otis-Lennon School Ability Test (OLSAT)	7 niveles K-12	Verbal, No Verbal, Grupos, Total	Harcourt Assessment	Stanford Achievement Test Metropolitan Achievement Tests
InView (antes, Test of Cognitive Skills)	6 niveles, 2-12	Verbal, No Verbal, Total	CTB/McGraw Hill	Terra Nova
Cognitive Abilities Test (CogAT)	13 niveles, K-12	Verbal, Cuantitativa, No Verbal, Compuesta	Riverside Publishing	Iowa Tests

Una de las características especiales de todas estas pruebas es su estructura multinivel. Podemos notar que la segunda columna del cuadro 9-2 indica el número de niveles. Recordemos que las pruebas de aplicación individual permiten cubrir un amplio rango de edades y capacidades gracias a las reglas de inicio y de discontinuación. Esa estrategia, evidentemente, no funcionará con una prueba de aplicación grupal en la que todos responden los mismos reactivos. Una estructura **multinivel** se adapta a las diferencias individuales –que son particularmente drásticas en edades jóvenes– por tener distintos niveles de la prueba para distintas edades o grados. Entonces, los diversos niveles están vinculados estadísticamente para permitir escalas continuas de puntuaciones a lo largo de todo el rango que cubre la prueba.

Prueba de Capacidad Escolar Otis-Lennon

La *Prueba de Capacidad Escolar Otis-Lennon* [Otis-Lennon School Ability], Octava Edición (OLSAT8; Otis & Lennon, 2003), es la versión más reciente de la larga línea de pruebas de inteligencia de Otis. Recordemos del capítulo 1 que el esfuerzo de Otis por crear una forma de aplicación grupal de una prueba tipo Binet representó un hito en la historia del campo de las pruebas. Las sucesivas ediciones de la prueba de Otis han estado entre las pruebas más usadas en el mundo durante casi 100 años. De acuerdo con el manual técnico del OLSAT8,

> el OLSAT8 está diseñado para medir las habilidades verbal, cuantitativa y de razonamiento figurativo que, con mayor claridad, están relacionadas con el aprovechamiento escolar. La serie OLSAT se basa en la idea de que, para aprender algo nuevo, el alumno debe ser capaz de percibir con agudeza, reconocer y recordar lo que ha percibido, pensar de manera lógica, comprender relaciones, abstraer a partir de lo concreto y hacer generalizaciones a contextos nuevos y diferentes. (Otis & Lennon, 2003, p. 5)

El OLSAT8 se usa principalmente en los programas de evaluación escolar. Las normas de esta prueba se obtuvieron junto con la *Prueba de Aprovechamiento Stanford*, Décima Edición.

Estructura y reactivos

El cuadro 9-3 bosqueja la estructura del OLSAT8, que es muy similar, aunque no idéntica, en los siete niveles de la serie. Un examinado responde uno de estos niveles.

Cuadro 9-3. Estructura del OLSAT

Comprensión verbal	Verbal	Total
Razonamiento verbal		
Razonamiento pictórico[a]	No Verbal	
Razonamiento figurativo		
Razonamiento cuantitativo[b]		

[a]Se encuentra sólo en los tres niveles más bajos (A, B, C).
[b]Se encuentra sólo en los tres niveles más altos (E, F, G).

La selección depende del grado escolar, edad y/o capacidad estimada del examinado. El cuadro 9-3 muestra los grados en que un nivel del OLSAT8 se usaría comúnmente; sin embargo, dado el modo en que la prueba se elaboró, es factible usar niveles atípicos. Un aplicador escolar toma la decisión respecto de qué nivel de la prueba emplear con un grupo de estudiantes; por ejemplo, un grupo muy brillante de segundo grado podría medirse de manera más eficaz con el nivel D que con el C, que sería más común para este grado, pues los estudiantes de este grupo podrían exceder el nivel C. La figura 9-1 ilustra cómo los diversos niveles cubren diferentes áreas del espectro de la capacidad. A la izquierda del espectro se encuentran los estudiantes de capacidad menor y/o más jóvenes, mientras que a la derecha están los de mayor capacidad y/o de grados superiores. Podemos notar que los niveles de la prueba se superponen; por ejemplo, los segmentos superiores del nivel A cubren la misma área que los segmentos inferiores del nivel B, y así sucesivamente. Este arreglo multinivel se aproxima al uso de las reglas de inicio y discontinuación de las pruebas de aplicación individual.

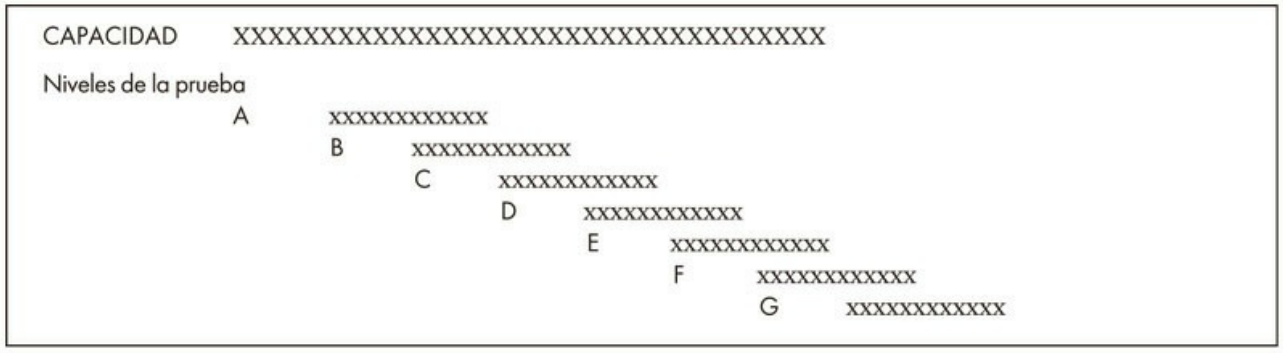


Figura 9-1. Ilustración de cómo una prueba multinivel cubre el espectro de la capacidad.

Ahora, podemos notar las categorías de los reactivos de la prueba. El OLSAT8 produce una puntuación total basada en todos los reactivos de cierto nivel (60-72 reactivos), subpuntuaciones verbal y no verbal basadas en 30-36 reactivos por área en cada nivel y puntuaciones de grupos. Estos últimos incluyen comprensión verbal y razonamiento verbal en el área verbal, y razonamiento pictórico, razonamiento figurativo y razonamiento cuantitativo en el área no verbal. Aún hay más subdivisiones de los reactivos en los grupos (p. ej., razonamiento oral y razonamiento aritmético dentro del grupo de razonamiento verbal), pero estas subdivisiones no producen ningún tipo de puntuación.

La provisión de varias puntuaciones diferentes es un desarrollo relativamente reciente en las series Otis. Las primeras ediciones producían sólo una puntuación global, aunque incorporaban varios tipos de reactivos. Podemos notar que el bosquejo del cuadro 9-3 ilustra un modelo jerárquico, como se describió en el capítulo 7.

Los reactivos de una prueba de capacidad mental de aplicación grupal aparecen en un cuadernillo de 12 a 20 páginas, como en el OLSAT8. Reactivos similares se encuentran

en la mayoría de las pruebas de este tipo. En la aplicación de la prueba, el examinador presentaría reactivos de práctica antes que los reactivos reales, así como las instrucciones de cómo marcar las respuestas, etc.

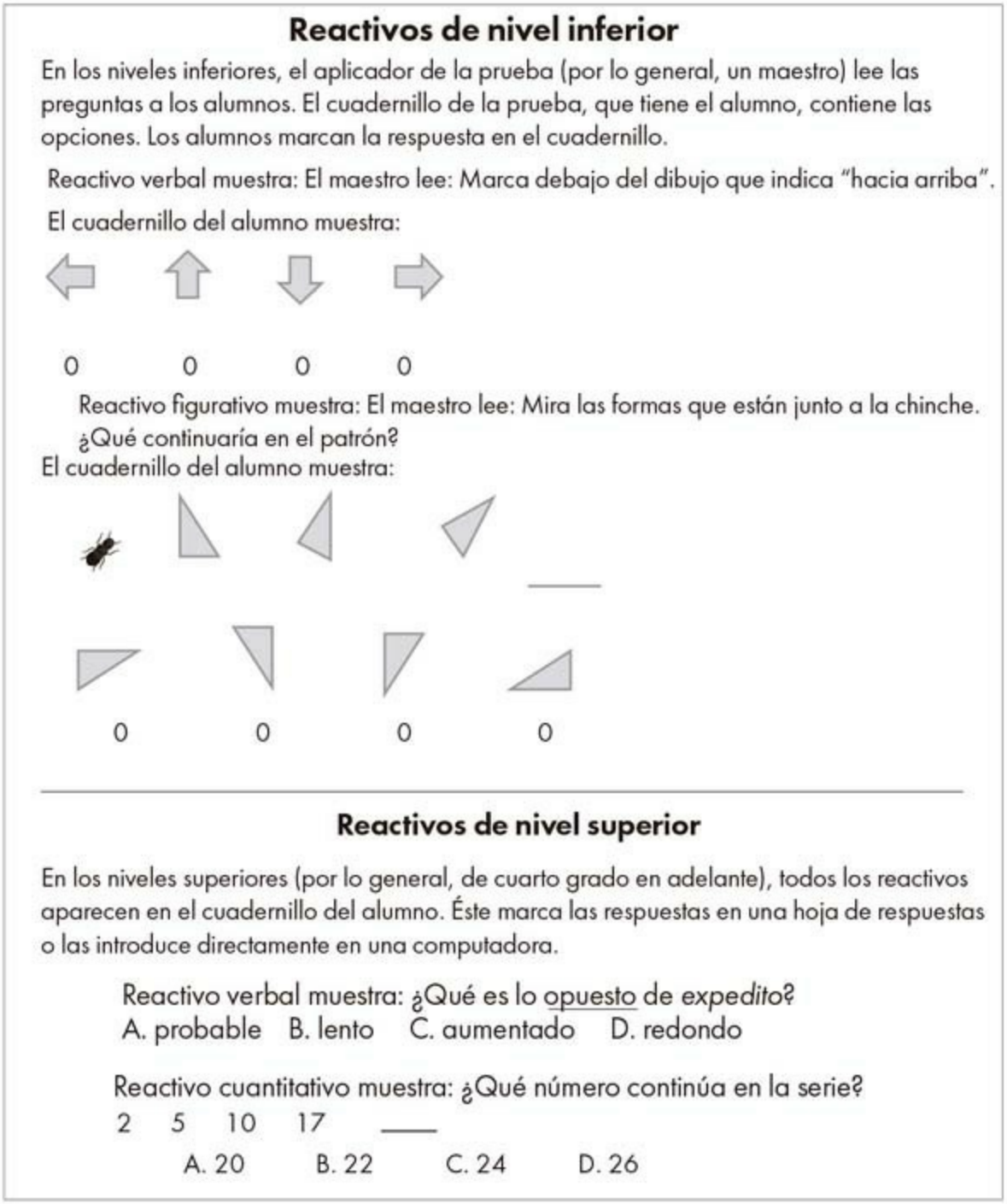


Figura 9-2. Reactivos muestra similares a los que aparecen en las pruebas grupales de capacidad mental.

¡Inténtalo!

Observa los reactivos de la figura 9-2. ¿Cómo podrías clasificar cada reactivo de acuerdo con las subcategorías de la columna izquierda del cuadro 9-3? Compara tus clasificaciones con las de alguien más. ¿Están de acuerdo?

¡Inténtalo!

Una auténtica industria artesanal se ha desarrollado para proveer de reactivos muestra y de práctica a las pruebas de capacidad mental que se abordan en este capítulo. Para ver ejemplos, sólo introduce el nombre de una prueba seguido de "practice items" [reactivos de práctica] en cualquier buscador de internet. Algunos sitios proporcionan ejemplos gratuitos, mientras que muchos otros sólo tratan de vender libros de reactivos muestra.

Puntuaciones, normas, estandarización

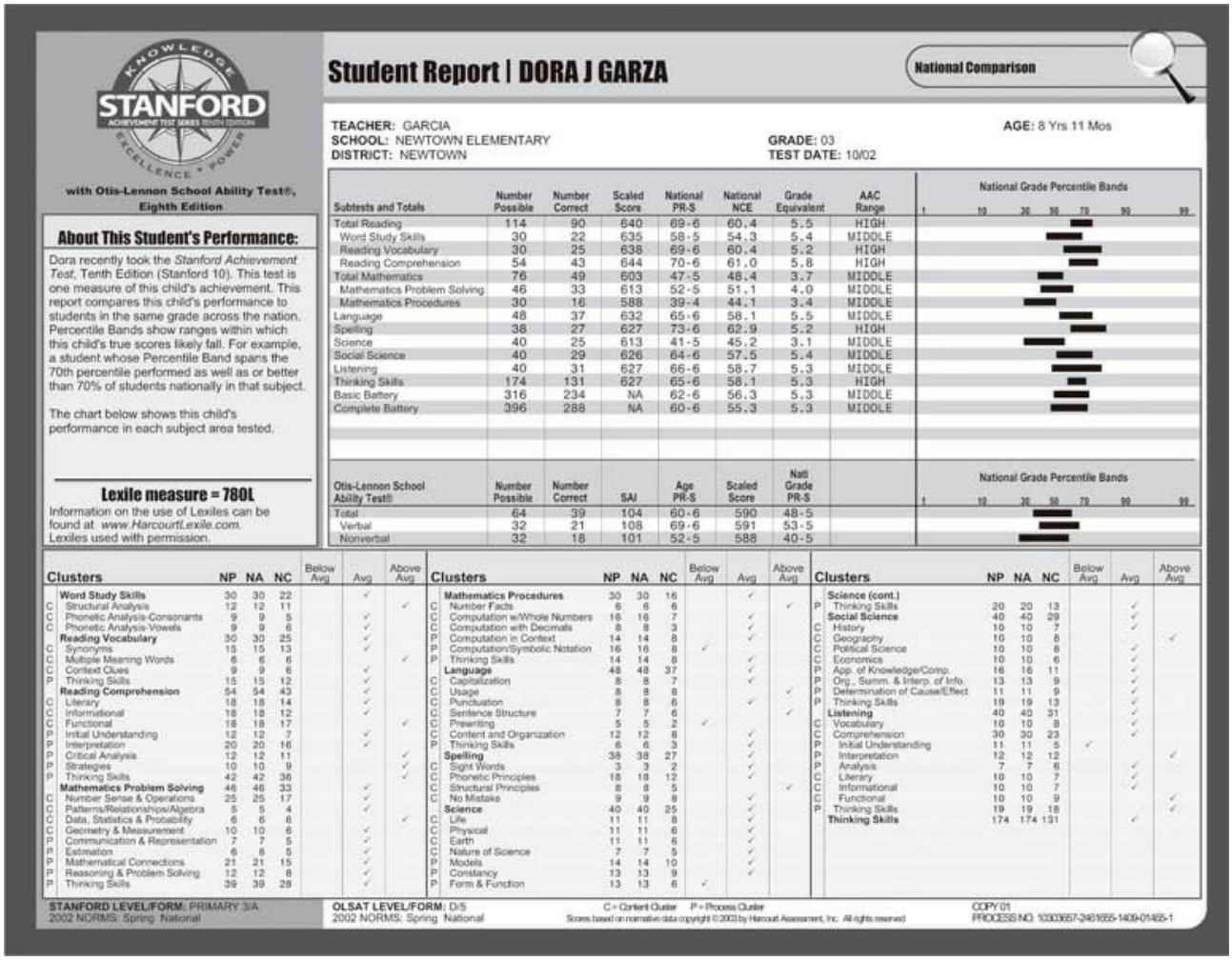
Como la mayoría de las pruebas, el OLSAT8 primero produce una puntuación natural o, en realidad, varias de ellas: Total, Verbal y No Verbal. Éstas se convierten en puntuaciones escalares, las cuales se usan primordialmente para convertir puntuaciones de todos los niveles en un sistema en común, pero rara vez se emplean para la interpretación práctica. Las puntuaciones escalares se convierten en el Índice de Capacidad Escolar (ICE). El ICE es una puntuación estándar con $M = 100$ y $DE = 16$; es el mismo sistema de puntuaciones que se usaba en el antiguo CI de las ediciones anteriores de las series Otis. Como señalamos antes en este capítulo, las editoriales de pruebas han hecho un esfuerzo para abandonar la terminología del CI, obviamente sin abandonarla por completo. Los ICE se determinan por separado para cada grupo de edad en intervalos de tres meses para edades de 5 a 19 años. Podemos notar que no hay normas separadas para los grupos adultos como en las pruebas de tipo Wechsler o el PPVT.

Los ICE pueden convertirse en rangos percentiles y estaninas, sea por grupo de edad o de grado. Los equivalentes de curva normal (ECN) se pueden derivar simplemente de los rangos percentiles. Las puntuaciones de grupos se convierten en una simple escala de tres categorías: debajo del promedio, promedio, arriba del promedio, que corresponden a las estaninas 1-3, 4-6 y 7-9, respectivamente.

En el uso típico del OLSAT8, una computadora realizaría todas estas conversiones y el usuario recibiría un informe impreso hecho por la computadora. Desde luego, también es posible hacer las conversiones (laboriosamente) a mano usando el cuadernillo de normas que forma parte de los materiales de la prueba. Las conversiones de los ICE en rangos percentiles, estaninas y ECN se realizan con ayuda de cuadros similares al cuadro 3-1.

La figura 9-3 presenta un informe muestra que contiene las puntuaciones del OLSAT8. Como se señaló antes, una prueba como el OLSAT8 se usa, por lo general, junto con

una prueba estandarizada de aprovechamiento. El informe de la figura 9-3 contiene puntuaciones del OLSAT8, así como de la *Prueba de Aprovechamiento de Stanford*, la cual examinaremos en el capítulo 11. Las puntuaciones del OLSAT8 aparecen en la sección central del informe e incluyen los ICE, rangos percentiles y estaninas. El desempeño en el OLSAT también está en el informe de otra manera, pues éste incorpora las *Anticipated Achievement Comparisons* [Comparaciones Anticipadas de Aprovechamiento] (AAC). El OLSAT8 se usa para predecir el desempeño en cada una de las pruebas Stanford. En la columna rotulada como AAC Range [Rango del AAC], el informe señala el resultado de esta comparación. "Alto" significa que la puntuación del aprovechamiento estuvo en el 23% superior de los casos para la estanina del OLSAT del estudiante. "Medio" significa que la puntuación del aprovechamiento estuvo en el 54% intermedio de los casos. "Bajo" significa que la puntuación del aprovechamiento estuvo en el 23% inferior de los casos. Esto se hace por separado con cada prueba de aprovechamiento. La figura 9-4 ilustra cómo funciona la comparación; desde luego, todos los cálculos se hacen "en la computadora" y en el informe sólo aparece el resultado. Todas las pruebas de capacidad mental enumeradas en el cuadro 9-2 proporcionan un informe que realiza, en esencia, estas mismas operaciones, aunque la metodología y terminología exactas pueden diferir del informe que se muestra aquí.



Student Report | DORA J GARZA

National Comparison

with Otis-Lennon School Ability Test®, Eighth Edition

TEACHER: GARCIA
SCHOOL: NEWTOWN ELEMENTARY
DISTRICT: NEWTOWN
GRADE: 03
TEST DATE: 10/02
AGE: 8 Yrs 11 Mos

About This Student's Performance:

Dora recently took the *Stanford Achievement Test*, Tenth Edition (Stanford 10). This test is one measure of this child's achievement. This report compares this child's performance to students in the same grade across the nation. Percentile Bands show ranges within which this child's true scores likely fall. For example, a student whose Percentile Band spans the 70th percentile performed as well as or better than 70% of students nationally in that subject.

The chart below shows this child's performance in each subject area tested.

Lexile measure = 780L

Information on the use of Lexiles can be found at www.HarcourtLexile.com. Lexiles used with permission.

Subtests and Totals	Number Possible	Number Correct	Scaled Score	National PR-S	National NCE	Grade Equivalent	AAC Range
Total Reading	114	90	640	69-6	60.4	5.5	HIGH
Word Study Skills	30	22	635	58-5	54.3	5.4	MIDDLE
Reading Vocabulary	30	25	638	69-6	60.4	5.2	HIGH
Reading Comprehension	54	43	644	70-6	61.0	5.8	HIGH
Total Mathematics	76	49	603	47-5	48.4	3.7	MIDDLE
Mathematics Problem Solving	46	33	613	52-5	51.1	4.0	MIDDLE
Mathematics Procedures	30	16	588	39-4	44.1	3.4	MIDDLE
Language	48	37	632	65-6	58.1	5.5	MIDDLE
Spelling	38	27	627	73-6	62.9	5.2	HIGH
Science	40	25	613	41-5	45.2	3.1	MIDDLE
Social Science	40	29	626	64-6	57.5	5.4	MIDDLE
Listening	40	31	627	66-6	58.7	5.3	MIDDLE
Thinking Skills	174	131	627	65-6	58.1	5.3	HIGH
Basic Battery	316	234	NA	62-6	56.3	5.3	MIDDLE
Complete Battery	396	288	NA	60-6	55.3	5.3	MIDDLE

National Grade Percentile Bands: 1 10 30 50 70 90 99

Otis-Lennon School Ability Test®	Number Possible	Number Correct	SAI	Age PR-S	Scaled Score	Nati Grade PR-S
Total	64	39	104	60-6	590	48-5
Verbal	32	21	108	69-6	591	53-5
Nonverbal	32	18	101	52-5	588	40-5

National Grade Percentile Bands: 1 10 30 50 70 90 99

	Clusters	NP	NA	NC	Below Avg	Avg	Above Avg
	Word Study Skills	30	30	22		✓	
C	Structural Analysis	12	12	11			✓
C	Phonetic Analysis-Consonants	9	9	5		✓	
C	Phonetic Analysis-Vowels	9	9	6		✓	
	Reading Vocabulary	30	30	25		✓	
C	Synonyms	15	15	13		✓	
C	Multiple Meaning Words	6	6	6			✓
C	Context Clues	9	9	6		✓	
P	Thinking Skills	15	15	12		✓	
	Reading Comprehension	54	54	43		✓	
C	Literary	18	18	14		✓	
C	Informational	18	18	12		✓	
C	Functional	18	18	17			✓
P	Initial Understanding	12	12	7		✓	
P	Interpretation	20	20	16		✓	
P	Critical Analysis	12	12	11			✓
P	Strategies	10	10	9			✓
P	Thinking Skills	42	42	36			✓
	Mathematics Problem Solving	46	46	33		✓	
C	Number Sense & Operations	25	25	17		✓	
C	Patterns/Relationships/Algebra	5	5	4		✓	
C	Data, Statistics & Probability	6	6	6			✓
C	Geometry & Measurement	10	10	6		✓	
P	Communication & Representation	7	7	5		✓	
P	Estimation	6	6	5		✓	
P	Mathematical Connections	21	21	15		✓	
P	Reasoning & Problem Solving	12	12	8		✓	
P	Thinking Skills	39	39	28		✓	

	Clusters	NP	NA	NC	Below Avg	Avg	Above Avg
	Mathematics Procedures	30	30	16		✓	
C	Number Facts	6	6	6			✓
C	Computation w/Whole Numbers	16	16	7		✓	
C	Computation with Decimals	8	8	3		✓	
P	Computation in Context	14	14	8		✓	
P	Computation/Symbolic Notation	16	16	8	✓		
P	Thinking Skills	14	14	8		✓	
	Language	48	48	37		✓	
C	Capitalization	8	8	7		✓	
C	Usage	8	8	8			✓
C	Punctuation	8	8	6		✓	
C	Sentence Structure	7	7	6			✓
C	Prewriting	5	5	2	✓		
C	Content and Organization	12	12	8		✓	
P	Thinking Skills	6	6	3		✓	
	Spelling	38	38	27		✓	
C	Sight Words	3	3	2		✓	
C	Phonetic Principles	18	18	12		✓	
C	Structural Principles	8	8	5			✓
C	No Mistake	9	9	8		✓	
	Science	40	40	25		✓	
C	Life	11	11	8		✓	
C	Physical	11	11	6		✓	
C	Earth	11	11	6		✓	
C	Nature of Science	7	7	5		✓	
P	Models	14	14	10		✓	
P	Constancy	13	13	9		✓	
P	Form & Function	13	13	6	✓		

	Clusters	NP	NA	NC	Below Avg	Avg	Above Avg
	Science (cont.)						
P	Thinking Skills	20	20	13		✓	
	Social Science	40	40	29		✓	
C	History	10	10	7		✓	
C	Geography	10	10	8			✓
C	Political Science	10	10	8		✓	
C	Economics	10	10	6		✓	
P	App. of Knowledge/Comp.	16	16	11		✓	
P	Org., Summ. & Interp. of Info.	13	13	9		✓	
P	Determination of Cause/Effect	11	11	9		✓	
P	Thinking Skills	19	19	13		✓	
	Listening	40	40	31		✓	
C	Vocabulary	10	10	8		✓	
C	Comprehension	30	30	23		✓	
P	Initial Understanding	11	11	5	✓		
P	Interpretation	12	12	12			✓
P	Analysis	7	7	6		✓	
C	Literary	10	10	7		✓	
C	Informational	10	10	7		✓	
C	Functional	10	10	9			✓
P	Thinking Skills	19	19	18			✓
	Thinking Skills	174	174	131		✓	

STANFORD LEVEL/FORM: PRIMARY 3/A
2002 NORMS: Spring National

OLSAT LEVEL/FORM: D/5
2002 NORMS: Spring National

C = Content Cluster P = Process Cluster

Scores based on normative data copyright ©2003 by Harcourt Assessment, Inc. All rights reserved.

COPY 01
PROCESS NO. 10303657-2461655-1409-01465-1

Figura 9-3. Informe muestra que incluye las puntuaciones del OLSAT8 y el AAC. Fuente: Stanford Achievement Test Series, 10a. ed. Copyright © 2003 por Harcourt Assessment, Inc. Reproducida con autorización. Todos los derechos reservados.

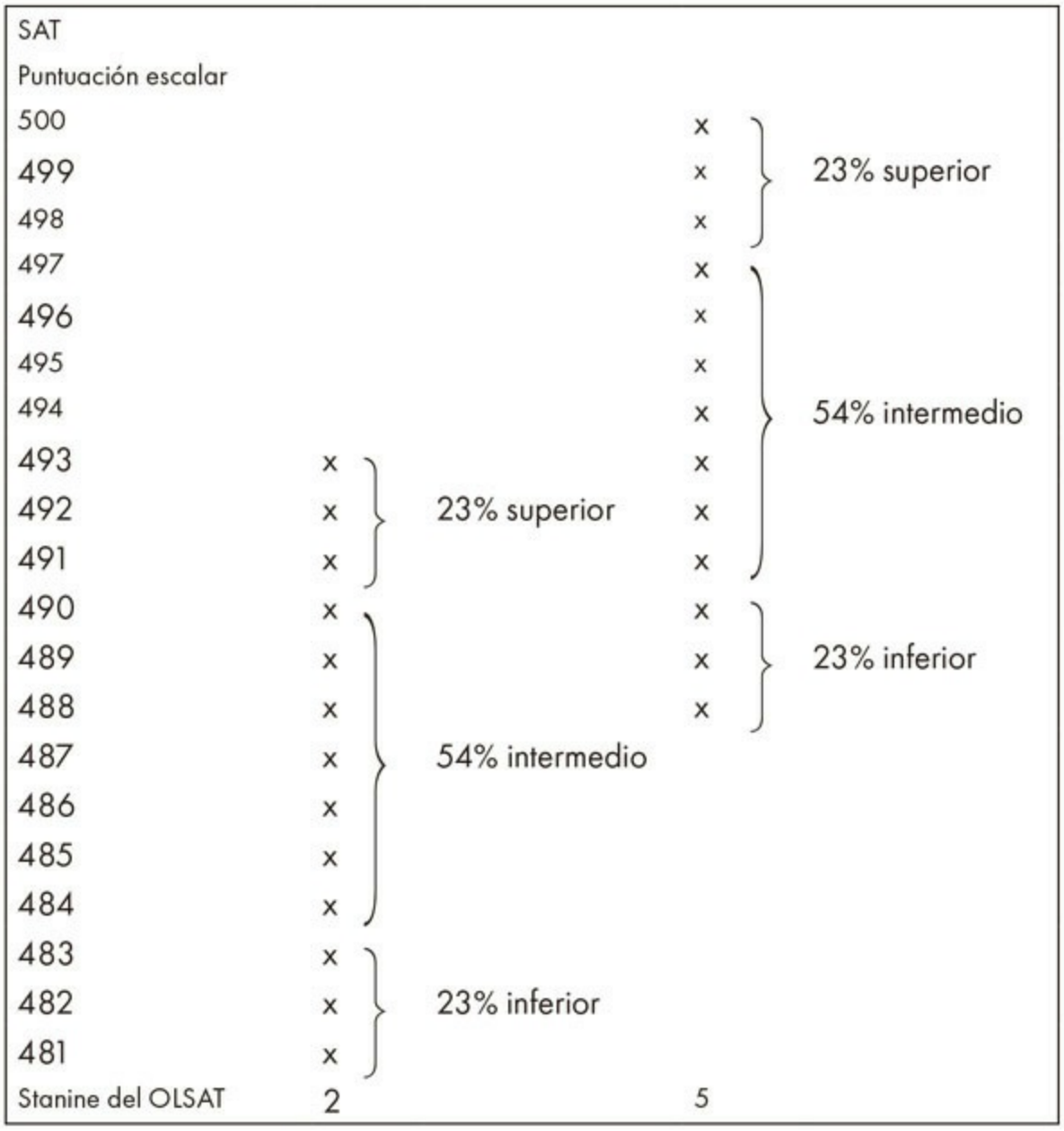


Figura 9-4. Metodología para obtener el AAC entre una prueba de capacidad mental y una de aprovechamiento (distribuciones ficticias).

El desarrollo de las normas del OLSAT8 demuestra la maestría que las principales editoriales de pruebas de la actualidad pueden poner en práctica, pues se han obtenido empíricamente para cada uno de los grados K-12 de manera separada para los períodos de primavera y otoño de cada año. Aproximadamente 275 500 estudiantes participaron en la estandarización de primavera y 135 000 en la de otoño. Si se suman los casos de los programas complementarios de investigación, casi medio millón de éstos contribuyeron a la investigación que fundamenta el OLSAT8.

¡Inténtalo!

En el caso que se muestra en la figura 9-3, ¿cuál es el ICE del alumno? ¿Aproximadamente a qué puntuación z corresponde esta puntuación?

En los grupos de normalización de primavera y otoño, los casos se estratificaron por nivel socioeconómico, región urbana/rural y etnia. También se balancearon los casos para mejorar su correspondencia con las características de la población total. El cuadro 9-4 resume la correspondencia entre el grupo de estandarización de otoño y la población nacional en relación con las variables de estratificación. Podemos notar que el ajuste entre los datos nacionales y los de la muestra de estandarización es muy cercano.

Cuadro 9-4. Comparación de la muestra de estandarización del OLSAT8 y los datos del censo nacional

Variable de estratificación	Nacional de EUA	OLSAT de estandarización
Región geográfica		
Noreste	19.7	17.0
Oeste medio	24.0	25.1
Sur	24.0	23.7
Oeste	32.3	34.9
Estatus socioeconómico		
Bajo	19.8	19.8
Bajo-medio	19.6	20.1
Medio	21.0	22.2
Alto-medio	19.7	18.6
Alto	19.9	19.3
Etnia		
Negro o afroamericano	15.1	15.0
Hispano o latino	15.0	14.3
Blanco	65.2	67.0
Asiático	3.8	3.5
Otros	0.9	0.2

Fuente: Otis, A. S., & Lennon, R. T. *Otis-Lennon School Ability Test*, 8a. ed. Copyright © 2003 por Harcourt Assessment, Inc. Reproducido con autorización. Todos los derechos reservados.

Los datos de estandarización de la mayoría de las pruebas grupales de capacidad mental y de las pruebas de aprovechamiento siguen patrones similares a lo que bosquejamos aquí acerca del OLSAT8. Desde los primeros días de los a veces azarosos o mal pensados procedimientos de estandarización, el campo ha alcanzado un alto nivel de sofisticación para determinar las normas.

Sin embargo, varios temas importantes aún nos llevan a hacer algunas advertencias.

Hay dos preocupaciones principales; primero, a menudo no sabemos qué porcentaje de estudiantes ha sido excluido de la muestra de estandarización por un sistema escolar, pues, sin duda, algunos de ellos lo fueron. El porcentaje quizá ronda el 5%. Más importante aún, por lo general sabemos poco acerca del nivel de motivación de los alumnos que participan en los programas de estandarización, lo cual puede ser particularmente problemático en los grados superiores, donde los alumnos suelen cooperar menos. Sin embargo, no sabemos mucho sobre este tema.

Confiabilidad

El manual técnico del OLSAT8 (Otis & Lennon, 2003) proporciona información sobre varios tipos de confiabilidad. La confiabilidad KR-20 que corresponde a las puntuaciones Totales en varios grados se ubica en una banda restringida que va de .89 a .94. Podemos usar .92 como un resumen apropiado de la consistencia interna dentro de un grado. La confiabilidad KR-20 de las puntuaciones Verbal y No Verbal va de .81 a .90, con una mediana de .84, en el caso Verbal, y de .86, en el No verbal. Podemos usar .85 como un resumen apropiado de la confiabilidad de la consistencia interna dentro de un grado de estas subpuntuaciones.

El manual del OLSAT8 también proporciona los coeficientes de confiabilidad KR-21 de la puntuación total, las subpuntuaciones Verbal y No Verbal y las puntuaciones de grupos. Estos datos aparecen por separado para cada grado de los períodos de primavera y otoño de estandarización. Es raro ver coeficientes de confiabilidad KR-21 en una prueba elaborada en años recientes, pues la fórmula KR-21 es una estimación que fue útil en la era anterior a la computadora, pero ahora es obsoleta. Más importante aún, los datos KR-21 muestran que las puntuaciones de grupos a menudo son muy poco confiables y probablemente no deberían aparecer como puntuaciones separadas. Aunque algunas de las confiabilidades de grupos son muy respetables, muchas se ubican alrededor de .55 y .65, claramente por debajo de los estándares del uso común.

Los cuadros completos de los datos de confiabilidad KR-20 y KR-21 también contienen los errores estándar de medición (EEM) de las puntuaciones. Los EEM aparecen en unidades de puntuación natural, lo cual es desafortunado, porque prácticamente nunca se interpretan las puntuaciones naturales de una prueba. Afortunadamente, el manual proporciona cuadros separados con los EEM de las puntuaciones Total, Verbal y No Verbal en unidades del Índice de Capacidad Escolar (ICE). Todos estos EEM están en el rango de 5.5 a 5.8, es decir, todos los EEM en unidades ICE son de cerca de un tercio de una desviación estándar (recordemos que en el OLSAT8 *DE* = 16). Ésta es una buena regla general: en el caso de una prueba bien construida y de extensión razonable, el EEM, por lo general, será de un tercio de la desviación estándar.

El manual del OLSAT8 no informa los datos de la confiabilidad de test-retest, lo cual es una omisión importante. Al menos, el manual debía informar los datos de test-retest de las versiones anteriores de la prueba. El manual argumenta en favor de la continuidad de

las ediciones sucesivas y presenta incluso correlaciones entre las ediciones séptima y octava.

Validez

El manual del OLSAT8 aborda la validez al hablar de contenido, relaciones con otras variables y estructura interna de la prueba. Asegura tener validez de contenido con base en la simple inspección de los reactivos, pero es más un asunto de validez aparente: ¿los reactivos tienen el aspecto de medir la capacidad relacionada con la escuela? El uso de la validez de contenido supone la existencia de un cuerpo bien definido de material contra el cual se puede contrastar el contenido de la prueba, pero no tenemos tal material en el caso de la capacidad escolar (o de la inteligencia o capacidad mental, los términos predecesores de las series Otis). El manual también argumenta que las correlaciones entre las ediciones séptima y octava demuestran la validez de contenido; sin embargo, es difícil seguir esta línea argumental.

En cuanto a las relaciones con otras variables, el manual incluye datos sobre las correlaciones del OLSAT y la *Prueba de Aprovechamiento de Stanford*, la cual revisaremos en detalle en el capítulo 11. El cuadro 9-5 presenta una selección de estas correlaciones. Por desgracia, el manual del OLSAT ofrece una mínima interpretación de estos datos.

Cuadro 9-5. Correlaciones entre OLSAT8 y SAT10

Grado	**Stanford**	**OLSAT8**		
		Total	**Verbal**	**No Verbal**
K	Total de Lectura	.68	.65	.61
	Total de Matemáticas	.73	.72	.62
3	Total de Lectura	.76	.78	.60
	Total de Matemáticas	.75	.72	.66
6	Total de Lectura	.69	.67	.63
	Total de Matemáticas	.73	.63	.73
9	Total de Lectura	.71	.70	.65
	Total de Matemáticas	.70	.61	.70

Fuente: Adaptado de Otis, A. S. y Lennon, R. T. *Otis-Lennon School Ability Test*, 8a. ed. (Cuadro 12).

Respecto a la estructura de la prueba, el manual presenta correlaciones entre niveles adyacentes, entre puntuaciones verbales y no verbales y correlaciones biseriales medianas. El razonamiento y el análisis que se usan en estas secciones del manual no están bien desarrollados. Además, considerando las afirmaciones de que se trata de una estructura jerárquica, es notable que no se presenten resultados del análisis factorial.

En Lennon (1985) se puede encontrar una descripción de las circunstancias que rodearon la elaboración de las pruebas Otis originales; incluso aparecen fragmentos de

una entrevista reveladora con Arthur Otis. Robertson (s.f.) hace un recuento del desarrollo de las primeras seis ediciones de las pruebas Otis.

Pruebas de admisión universitarias

La segunda categoría importante de las pruebas de capacidad mental de aplicación grupal son las de admisión a la universidad. Esta categoría, al menos en EUA, está formada sólo por dos pruebas: el SAT y el ACT. El propósito principal de estas pruebas es auxiliar en la selección de estudiantes para ingresar a la universidad. Los aspirantes llegan con una gran diversidad de experiencias en el bachillerato. Prácticas de valoración escolar y experiencias estudiantiles individuales en cursos, demandas laborales y obligaciones extracurriculares pueden diferir de manera considerable. Las pruebas de selección universitaria ofrecen un criterio uniforme para medir las capacidades de los alumnos y sus posibilidades de éxito en la universidad. Las dos pruebas que se emplean en la actualidad abordan el proceso de manera diferente en cierto modo, pero sus propósitos fundamentales son muy similares. Estas pruebas también se pueden usar para ubicar a los estudiantes en cursos universitarios una vez que han sido admitidos.

SAT

Pocas pruebas igualan al SAT en reconocimiento y publicidad de su nombre. Su historia se remonta al establecimiento del campo de las pruebas como disciplina diferenciada. Cada año, millones de estudiantes que terminan el bachillerato (y sus padres) se preparan con ansiedad para esta prueba y luego esperan los resultados. Los periódicos están llenos de informes sobre las fluctuaciones de los promedios estatales y nacionales del SAT. Se trata de un ejemplo clásico de una prueba atrevida, más de lo necesario dada la manera en que las universidades usan en realidad las pruebas, pero atrevida aun así.

Empecemos la descripción del SAT haciendo algunas distinciones entre los términos. Más que en el caso de cualquier otra prueba, las personas se sienten confundidas con los términos que rodean al SAT. Examinemos las fuentes de esta confusión para asegurarnos de que hacemos las distinciones apropiadas en relación con el nombre del SAT, el College Board y el ETS.

El SAT es una *prueba*. Desde sus orígenes y hasta 1988, el SAT fue acrónimo de *Scholastic Aptitud Test* [Prueba de Aptitud Escolar] y, después, de *Scholastic Assessment Test* [Prueba de Evaluación Escolar]. Muchas fuentes y referencias populares aún usan estos nombres. La sustitución de “aptitud” por “evaluación” fue resultado del esfuerzo de la profesión para evitar la implicación de que los rasgos que se miden son, de alguna manera, innatos.

El SAT es un grupo de pruebas. Primero que nada, existe el SAT (antes *SAT Reasoning Test* [Prueba de Razonamiento SAT]), que en realidad es tres pruebas, como se describe más adelante. Ésta es la muy usada prueba de admisión a la universidad, a la que la gente se refiere cuando habla del SAT. Segundo, está el *SAT Subject Test* [Prueba por Temas SAT], una serie de pruebas separadas en campos como biología, historia, química y numerosos idiomas. Para complicar aún más las cosas respecto del nombre del

SAT, la batería estandarizada de aprovechamiento más usada es el *Stanford Achievement Test* [Prueba de Aprovechamiento Stanford], también SAT. Así, las iniciales SAT a veces se refieren a esta prueba de aprovechamiento, que no tiene ninguna relación con la prueba de admisión a la universidad SAT. Sólo se requiere estar atento a la posible confusión entre estas dos SAT.

College Board es *una asociación* de cerca de 3000 universidades y escuelas de bachillerato que se enfocan en la transición del bachillerato a la universidad. Esta asociación, cuya sede se encuentra en la ciudad de Nueva York, patrocina y supervisa el desarrollo del SAT, así como muchos otros programas de evaluación. Originalmente, el nombre de la asociación fue College Entrance Examination Board [Consejo de Exámenes de Entrada a la Universidad] o CEEB, acrónimo que aún se encuentra con frecuencia. El CEEB se organizó en 1900 con el fin específico de coordinar la evaluación de ingreso a la universidad entre las escuelas más selectivas, las cuales solían aplicar sus propios programas de evaluación para elegir entre los aspirantes (Johnson, 1994). El SAT fue el resultado de la coordinación de estos esfuerzos y se aplicó por primera vez en 1901. La prueba tuvo forma de ensayo en su totalidad hasta 1926, cuando adoptó el formato de opción múltiple. Debido a que College Board patrocina la elaboración del SAT, mucha gente lo llama "el College Boards".

ETS significa Educational Testing Service [Servicio de Evaluación Educativa], que es una *organización* sin fines de lucro para la *elaboración de pruebas* con sede en Princeton, Nueva Jersey. El ETS, en realidad, elabora el SAT y organiza su extensa aplicación. El ETS hace esto por un convenio con el College Board; el ETS elabora y vende muchas otras pruebas.

Estructura y reactivos

Durante muchos años, las publicaciones del College Board hicieron hincapié en que el SAT intenta medir capacidades muy generalizadas desarrolladas en un largo periodo; es decir, las pruebas no eran medidas de "aptitud innata". En años recientes, se han hecho algunos intentos para mostrar que el contenido de la prueba se basa en las habilidades que se deben desarrollar en la escuela, lo que la hace parecer más una prueba de aprovechamiento que de capacidades generalizadas que trascienden el aprendizaje escolar específico (Lawrence *et al*., 2002).

El SAT consta de tres pruebas: Lectura crítica, Matemáticas y Escritura. Esta estructura de tres pruebas se introdujo en 2005, pues antes estaba formado por dos pruebas: Verbal y Matemáticas. Muchas fuentes siguen refiriéndose al SAT Verbal y al SAT de Matemáticas. El cuadro 9-6 bosqueja el número y tipo de reactivos en las tres pruebas del SAT. Además, hay una sección experimental empleada para la investigación y propósitos de desarrollo. Los reactivos experimentales no contribuyen a la puntuación del alumno.

Cuadro 9-6. Estructura del SAT

Área	Tipo de reactivos	Tiempo (min)	
		Secciones	Total
Lectura crítica	Frases incompletas Lectura de pasajes	25, 25, 20	70
Matemáticas	Opción múltiple Respuesta producida por el alumno	25, 25, 20	70
Escritura	Opción múltiple Ensayo	35 25	60
Experimental (no se califica)	Opción múltiple	25	25
Total			200 (3 hrs, 45 min)[a]

[a] El tiempo total incluye el de la sección para familiarizarse, pero no para distribuir los materiales, leer las instrucciones ni otras cuestiones de aplicación.

Tradicionalmente, todos los reactivos del SAT han sido de opción múltiple, pero en años recientes, en la prueba Matemáticas también se han incluido algunos **reactivos de llenar óvalos** por parte del alumno, en los que éste determina una respuesta numérica (p. ej., 524) y marca la respuesta rellenando el óvalo correspondiente. El SAT más reciente ha eliminado dos tipos de reactivos que fueron "famosos" en las primeras ediciones: analogías verbales y comparaciones cuantitativas. Y, desde luego, la reciente edición tiene la prueba Escritura, que es por completo nueva e incluye un ensayo que es calificado por dos jueces "humanos" mediante un método holístico, como se describió en el capítulo 6. El SAT usa una corrección por adivinación (véase capítulo 3) en el caso de reactivos de opción múltiple.

Puntuación escalar y normas

Junto con la escala del CI, la escala del SAT es una de las más conocidas en el mundo de la psicometría. El sistema de puntuación escalar para cada una de las pruebas varía entre 200 y 800, con $M = 500$ y $DE = 100$. La situación no es tan sencilla, pero estos valores generalizados son puntos de referencia útiles. También se proporcionan los rangos percentiles nacionales de cada prueba del SAT. Tanto las puntuaciones escalares como los percentiles son normas del usuario, es decir, se basan en quien responde el SAT, aunque no necesariamente sea representativo de una población bien definida.

En la anterior estructura de dos pruebas (Verbal y Matemáticas), cada una con su escala de 200 a 800, se acostumbraba combinar las puntuaciones en un SAT Total. Ésta no era una puntuación oficial del SAT, pero la práctica de usarlo se extendió. Así, un estudiante con un "Boards de 1500" era un trofeo intelectual que, tal vez, tenía puntuaciones en el rango de 700 a 800 en las pruebas del SAT (Verbal y Matemáticas). Con la nueva estructura de tres pruebas, aún no es claro cómo las personas podrían combinar las puntuaciones. El "Boards de 1500" basado en tres pruebas, cada una con

un sistema de puntuaciones estándar (escalares) con M = 500 y DE = 100, indica de manera perfecta el promedio más que el desempeño superior. La situación ilustra la importancia de saber con exactitud qué pruebas y qué sistemas de puntuación se emplean en la interpretación del desempeño en la prueba.

Las normas de puntuaciones escalares y de rangos percentiles del SAT tienen diferentes bases. Las primeras se determinaron con un grupo nacional en 1994. Conforme se han aplicado más pruebas cada año, las nuevas se igualan e informan en términos de estas normas nacionales. De ahí que el promedio nacional real pueda fluctuar en el sistema de puntuación escalar entre 200 y 800. Esto ofrece las bases para nuevos informes sobre las puntuaciones del SAT que aumentan o disminuyen. Por otro lado, las normas de percentiles del SAT se ajustan de manera continua cada año, pues surgen nuevas normas de percentiles con los examinados que responden la prueba cada año. Esta combinación de una norma de puntuación escalar fijada en un momento y normas de percentiles que cambian cada año lleva a una extraña circunstancia: nuestras reglas acerca de las relaciones entre los sistemas de puntuación (véase cuadro 3-1 y figuras 3-9 y 3-10) no se cumplen. Por ejemplo, una puntuación escalar de 500 no necesariamente se ubica en el percentil 50, lo cual requiere hacer malabares psicométricos para explicarlo.

ACT

Al igual que con el SAT, comenzaremos con algunas aclaraciones. Las iniciales ACT corresponden, por lo general, a *American College Test*, pero designan tanto una organización como una prueba. La organización es ACT, Inc., con sede en la ciudad de Iowa, Iowa. La prueba ahora se conoce sólo por las iniciales ACT, antes *ACT Assessment* [Evaluación ACT]. ACT, Inc. crea y distribuye la prueba; también elabora otras pruebas y financia programas de investigación, casi siempre en niveles universitarios y preuniversitarios. El *ACT* (*Assessment*) es la empresa emblemática de ACT, Inc. Técnicamente, el *ACT* incluye pruebas académicas y diversas fuentes auxiliares de información: un perfil de las características y antecedentes del estudiante, un autorreporte de los cursos y grados del bachillerato y un breve inventario de intereses. Hablar del ACT suele implicar pruebas académicas más que el paquete completo con los materiales auxiliares.

Estructura y reactivos

El método del ACT es algo diferente del enfoque del SAT. Mientras que el SAT ha hecho hincapié tradicionalmente en capacidades muy generalizadas (aunque, como señalamos antes, el énfasis ha ido cambiando), el ACT siempre ha resaltado la evaluación de habilidades escolares. En su forma original, el ACT producía puntuaciones en inglés, matemáticas, estudios sociales y ciencias. En esencia se trató de una extensión hacia arriba del *Iowa Test of Educational Development* [Prueba Iowa de Desarrollo

Educativo], una prueba tradicional de aprovechamiento para las escuelas secundarias. Hoy, las pruebas son ligeramente distintas; sin embargo, el enfoque básico sigue siendo el mismo. Hay un énfasis en la base curricular de la escuela en la prueba.

Dada esta orientación, podríamos preguntar: ¿por qué clasificamos esta prueba como de capacidad mental en vez de aprovechamiento? Es una buena y legítima pregunta. El factor determinante para clasificar el ACT como prueba de capacidad mental es su uso principal, que es, principalmente, predecir el desempeño futuro en la universidad. Su validez depende sobre todo de qué tan bien hace estas predicciones. Esa es la razón por la que la clasificamos como prueba de capacidad mental.

El cuadro 9-7 bosqueja la estructura del ACT. Éstas son las cuatro pruebas principales: Inglés, Matemáticas, Lectura y Razonamiento científico. Tres de estas pruebas tienen subpruebas; también hay una puntuación Compuesta, la cual consiste en el promedio de las cuatro pruebas. La mayoría de las veces que se habla del ACT, nos referimos a la puntuación Compuesta; por ejemplo, si una universidad afirma que la admisión normal requiere un ACT de 22, esto se refiere a la puntuación Compuesta. En 2005, el ACT agregó la prueba Escritura, la cual es opcional y no forma parte de la puntuación Compuesta.

Cuadro 9-7. Estructura del ACT

Área	Subpruebas	Reactivos	Tiempo (min)
Inglés	Uso/Mecánica Habilidades retóricas	75	45
Matemáticas	Preálgebra/Álgebra elemental Álgebra intermedia/Geometría espacial Geometría plana/Trigonometría	60	60
Lectura Ciencia	Estudios sociales/Ciencias Artes/Literatura (no subpruebas)	40 40	35 35
COMPUESTO		215	2 hrs, 55 min
Prueba escrita (opcional)		-	30

¡Inténtalo!

Para ver ejemplos de los temas que se emplean para los ensayos de las partes escritas del SAT y ACT, incluyendo ejemplos de las rúbricas de calificación, entra a los siguientes sitios de internet:
ACT: www.actstudent.org/writing/sample/index.html
SAT: www.collegeboard.com/student/testing/sat/about/sat/writing.html\TIL\

Escala de puntuaciones y normas

Las puntuaciones de las cuatro pruebas del ACT se ubican en una escala de 1 a 36, la

cual es, sin duda, una de las más inusuales del mundo de la psicometría. Se define a sí misma con un rango más que con una media y una desviación estándar. La inspección de los cuadros de normas del ACT revela que la escala tiene una media aproximada de 20 y una desviación estándar aproximada de 5, aunque estos valores fluctúan ligeramente de una prueba a otra.

Sin embargo, es útil recordar los valores de $M = 20$ y $DE = 5$ a pesar de no ser la base de la escala.

La puntuación Compuesta es el simple promedio de las puntuaciones de las cuatro pruebas redondeado al número entero más cercano. La media y la desviación estándar del Compuesto son más o menos las mismas que las de las cuatro pruebas: $M = 20$ y $DE = 5$. Las subpruebas también tienen una escala inusual de puntuaciones, que va de 1 a 18. La inspección de los cuadros de las normas revela que las medias y desviaciones estándar de estas escalas rondan el 10 y el 3, respectivamente.

Las normas del ACT son simples normas del usuario, es decir, se basan en cualquiera que responda la prueba en un ciclo anual, por lo general, cerca de un millón de estudiantes.

Generalizaciones acerca de la confiabilidad de las puntuaciones del SAT y el ACT

Se ha estudiado ampliamente la confiabilidad de las puntuaciones tradicionales del SAT y el ACT, pero no los nuevos segmentos para escribir ensayos. Los datos de confiabilidad de las dos series de pruebas son lo suficientemente similares tal que podemos presentar un solo resumen. El cuadro 9-8 muestra los resultados típicos de los estudios de confiabilidad de estas pruebas (véase ACT, 2007; Breland *et al*., 2004; College Board, 2012; Ewing *et al*., 2005). Los resultados típicos son predecibles en su mayor parte a partir del número de reactivos que forman parte de una puntuación particular. Las puntuaciones totales, basadas en más de 150 reactivos, tienen una confiabilidad alta, por lo general, de alrededor de .95. Las pruebas principales, casi siempre basadas en al menos 40 reactivos, tienden a tener confiabilidades no menores de .85 y, a menudo, ligeramente superiores a .90. Las puntuaciones totales y las pruebas principales son los vehículos más importantes para los informes del SAT y el ACT; en algunas aplicaciones se hace el intento de proporcionar subpuntuaciones a un nivel por debajo de las pruebas principales, pero, en varios casos, los niveles de confiabilidad se encuentran por debajo de lo deseado. Las confiabilidades se pueden predecir, en gran parte, a partir del número de reactivos del grupo. Luego está el caso de las pruebas escritas. Como señalamos antes, no se han estudiado de manera sistemática las confiabilidades de estas pruebas, al menos la del ensayo. Sin embargo, los datos disponibles en la actualidad muestran que las confiabilidades interjueces son bastante buenas, mientras que las correlaciones con temas alternos son problemáticas.

Cuadro 9-8. Resultados típicos de los estudios de confiabilidad del SAT y el ACT

Tipos de puntuaciones	**r**
Puntuaciones totales (p. ej., Compuesto del ACT, SAT-CR + SAT-M)	.95
Pruebas principales (p. ej., Lectura, Matemáticas del ACT, SAT-CR, SAT-M)	.85-.90
Subpuntuaciones (grupos dentro de las pruebas principales)	.65-.85
Escritura (sección para escribir un ensayo)	formas alternas .65-.70, interjueces .80-.95

Generalizaciones acerca de la validez del SAT y el ACT

El método principal para estudiar la validez del SAT y el ACT es la validez predictiva. Literalmente se han realizado miles de estudios de este tipo de validez con estas pruebas; aquí presentamos un resumen de los resultados típicos de estos estudios. Los resultados del SAT y el ACT son suficientemente similares para poder hacer un solo resumen, que se basa primordialmente en Burton y Ramist (2001), Bridgeman, McCamley-Jenkins y Ervin (2000) y Kobrin *et al*. (2008), en lo que respecta al SAT, y en *ACT Assessment Technical Manual* (ACT, 2007), en lo tocante al ACT. En la actualidad, no hay suficiente evidencia respecto de la contribución de las pruebas escritas a la validez predictiva que nos permita hacer generalizaciones claras, por lo que no tratamos aquí dichas pruebas.

El diseño de los estudios, por lo general, sigue este patrón. El criterio más común (que se intenta predecir) es el *GPA del primer año en la universidad*, denominado FYGPA (siglas de first-year college GPA), y los predictores son las pruebas. En los estudios suelen representarse las pruebas por separado (Verbal y Matemáticas del SAT, y las cuatro pruebas del ACT), así como en términos de puntuaciones totales, pero nosotros nos concentraremos en las puntuaciones totales y nos referiremos al SAT y ACT simplemente como *pruebas de admisión* (PA). Sin excepción, los estudios incluyen las notas del bachillerato, que pueden estar representadas por el GPA de bachillerato, tomarse de los expedientes de la escuela o ser informadas por el propio alumno; también pueden estar representadas por un GPA de bachillerato calculado, por ejemplo, con base sólo en los cursos pertinentes para los estudios universitarios, o por un rango de bachillerato. Emplearemos RB para designar el registro de bachillerato sin importar la manera en que se determine.

- Los estudios típicos informan lo siguiente:
- Correlación entre la prueba de admisión (PA) y el FYGPA
- Correlación entre el registro de bachillerato (RB) y el FYGPA
- Correlación múltiple de (PA + BR) y el FYGPA

El efecto de la **restricción de rango** en las correlaciones resultantes es un tema de especial preocupación en estos estudios. Éstos se llevan a cabo con estudiantes que ya han sido admitidos y, por lo tanto, suelen tener menor variabilidad que el conjunto entero

de aspirantes a la universidad. En el capítulo 4, revisamos la naturaleza de este problema y cómo hacer correcciones al respecto. Los estudios de validez predictiva de las pruebas de admisión a la universidad, por lo general, informan las correlaciones sin corrección y corregidas. La imperfecta confiabilidad del criterio (FYGPA) también es tema de preocupación, pues algunos estudios introducen correcciones relacionadas con ella.

Cuadro 9-9. Resultados típicos de los estudios de validez predictiva de las pruebas de admisión a la universidad

	Correlaciones		
Variables	**Sin corrección**	**Corregidas por rango**	**Corregidas por rango y confiabilidad**
FYGPA–PA	.40	.50	.55
FYGPA–RB	.40	.50	.55
FYGPA–(PA + RB)	.50	.60	.65

El cuadro 9-9 presenta un resumen de los resultados típicos de numerosos estudios de este tipo. Las correlaciones corregidas de acuerdo con el rango aparecen en negritas, porque son probablemente las estimaciones más realistas de cómo funcionan estas pruebas. Hacemos las siguientes generalizaciones:

1. PA y RB tienen las mismas correlaciones con el FYGPA.

2. La combinación de PA y RB siempre es mejor que cualquiera de los dos por separado para predecir el FYGPA. Esta combinación emplea la metodología de la correlación múltiple, por lo que no se trata de la simple suma de dos variables. La validez incremental de la combinación es casi de .10.

3. El examen de los datos originales indica que las correlaciones que aparecen en el cuadro 9-9 son aproximadamente las mismas en los distintos grupos raciales/étnicos, pero hay algunas diferencias por género, carrera universitaria y tipo de universidad. Sin embargo, las semejanzas, a lo largo de estas diversas categorías, son más impresionantes que las diferencias. No obstante, para usos prácticos, cada universidad debe llevar a cabo sus propios estudios de validez.

¿Qué concluimos acerca de la utilidad de las pruebas de admisión a la universidad a partir de este resumen? Quienes están a favor del uso de las pruebas subrayan que una correlación de .50 es un coeficiente de validez muy respetable. De hecho, es notable que los resultados de unas pocas horas de evaluación tengan una correlación así de alta con el FYGPA que pueda resumir un año entero de esfuerzo. Además, una PA casi siempre añade poder predictivo al índice del RB. A quienes no les gusta la idea de aplicar pruebas de admisión a la universidad señalan que una correlación de .50 explica sólo 25% de la varianza del FYGPA, lo cual deja 75% a otros factores. Del mismo modo, los oponentes argumentan que la validez incremental de la prueba de admisión, que añade .10 al poder predictivo del índice de RB, no compensa los problemas asociados con esta arriesgada evaluación. Tal vez estos argumentos, a favor y en contra, continúen para siempre.

Selección de graduados y profesionales

Una tercera aplicación importante de las pruebas de capacidad mental de aplicación grupal es para la selección de estudiantes graduados y profesionales para programas de posgrado o profesionalización. La preparación de un estudiante universitario (como lo indica el GPA de estudiantes universitarios), su experiencia en el campo, su motivación para continuar con estudios de posgrado (como se documenta en cartas de recomendación) y una gran variedad de otros factores ayudarán a elegir a los estudiantes más promisorios. Sin embargo, hay un sentimiento generalizado de que, más allá de estos factores, la capacidad mental general es importante, por lo que las pruebas de esta categoría buscan evaluar esta capacidad para estos fines. El cuadro 9-10 enumera algunas de las principales pruebas de esta categoría. Nosotros ilustraremos esta aplicación describiendo sólo un ejemplo, el *Graduate Record Examination: General Test* [Examen de Registros de Graduados: Prueba General]. Sin embargo, las observaciones sobre esta prueba, por ejemplo, respecto de su confiabilidad y validez, se aplican sorprendentemente bien a la mayoría de las pruebas que aparecen en el cuadro 9-10. Debemos hacer notar que algunas de estas pruebas estarían mejor clasificadas, al menos en parte, como pruebas de aprovechamiento más de de capacidad general; por ejemplo, el MCAT y el DAT abordan una mezcla de conocimientos sobre distintos temas (p. ej., fisiología, física) y capacidades más generales (p. ej., capacidad verbal). Para obtener más información de las pruebas que aparecen en el cuadro 9-10, revisa sus sitios web, así como las fuentes usuales como el *Mental Measurements Yearbook* de Buro.

Cuadro 9-10. Ejemplos de pruebas usadas para la selección en programas de posgrado y profesionalización

Acrónimo	Prueba	Aspirantes a:	Sitio web
GMAT	Graduate Management Admission Test	Escuelas de negocios, en especial el MBA	www.gmac.com/gmat
GRE	Graduate Record Examinations	Programas de posgrado, en especial doctorales	www.gre.org
LSAT	Law School Admission Test	Escuela de derecho	www.lsac.org
MCAT	Medical College Admission Test	Escuela de medicina, en especial MD, DO	www.aamc.org/students/mcat
MAT	Miller Analogies Test	Escuela de posgrado	www.milleranalogies.com
DAT	Dental Admission Test	Escuela odontológica	www.ada.org/prof/dat.aspx

¡Inténtalo!

Entra a los sitios web citados en el cuadro 9-10 y revisa la información sobre las pruebas. ¿Cuál es el

propósito de cada una?

Graduate Record Examinations: Prueba General

El **Graduate Record Examinations (GRE)**[1] abarca las pruebas General y Temas. Entre las de Temas se incluyen ocho pruebas de aprovechamiento basadas en el contenido típico de carreras universitarias, como biología, matemáticas o psicología. En el siguiente capítulo, discutiremos las pruebas de aprovechamiento. El GRE Prueba General, que describimos aquí, constituye un buen ejemplo de una prueba de capacidad mental general diseñada para público y propósitos muy específicos: ayudar en la selección de aspirantes a programas de posgrado, en especial de doctorado.

Estructura y reactivos

El GRE Prueba General consta, en la actualidad, de tres pruebas separadas, Razonamiento verbal, Razonamiento cuantitativo y Escritura analítica, que se suelen representar como GRE-V, GRE-Q [Quantitative] y GRE-AW [Analytical Writing], respectivamente.[2] La figura 9-5 bosqueja la estructura de las pruebas actuales (ETS, 2012a, 2012b).

Razonamiento verbal (dos secciones de 30 min y 20 reactivos)	**Razonamiento cuantitativo** (dos secciones de 35 min y 20 reactivos)
Comprensión de lectura	Ecuaciones y expresiones matemáticas
Completamiento de texto	Comparaciones cuantitativas
Equivalencia de secuencias	Conjuntos de interpretaciones de datos
Escritura analítica (dos ensayos de 30 min cada uno)	
Analizar un tema	
Analizar un argumento	

Figura 9-5. Esquema de la estructura del GRE Prueba General revisada.

Durante muchos años, el GRE-G empleó un formato tradicional de lápiz y papel y extensión fija, pero ahora la mayoría de los individuos responde un formato adaptable por computadora en un centro de evaluación Prometric; la versión en formato tradicional está disponible para usarse cuando no esté disponible un centro de evaluación. El bosquejo de la figura 9-5 es de una versión adaptable mediante computadora por sección, no por reactivo, es decir, se responde una sección entera y luego se pasa a una más o menos difícil. En el sistema más conocido que es adaptable reactivo por reactivo, el individuo es "dirigido" hacia arriba o hacia abajo después de cada reactivo. El GRE

Prueba por Temas sigue usando el tradicional formato de lápiz y papel, aunque se están desarrollando versiones adaptadas para computadora.

La mayoría de los reactivos del GRE Prueba General revisada sigue siendo de opción múltiple (OM) y cinco opciones con una sola respuesta correcta. Sin embargo, la prueba introduce algunos formatos de reactivos inusuales; por ejemplo, algunos reactivos presentan varias opciones (es decir, OM común), pero dos, tres o incluso todas las opciones son correctas. En estos reactivos se debe elegir todas las opciones correctas para obtener el crédito, pues no hay crédito parcial. La figura 9-6 muestra un ejemplo muy sencillo de estos reactivos.

¿Cuál es la capital del estado de Nueva York?
(Elige una de las respuestas.)
A. Albania
B. Brooklyn
C. Buffalo
D. Ciudad de Nueva York
E. Trenton

¿Cuáles de éstas son capitales de estado?
(Elige todas las que lo sean.)
A. Albania, Nueva York
B. Chicago, Illinois
C. Columbus, Ohio
D. Juneau, Alaska
E. Dallas, Texas
F. Scranton, Pennsylvania

Figura 9-6. Ejemplo sencillo de reactivos de opción múltiple de una sola respuesta correcta y de más de una respuesta correcta.

En otros reactivos se hace clic en una oración de un párrafo para elegir una respuesta. En la prueba Cuantitativa, en algunos reactivos, en realidad se introducen números en una casilla para indicar la respuesta.

En la prueba de Escritura analítica, dos jueces califican cada ensayo con una rúbrica holística de seis puntos; si la valoración de los dos jueces difiere por más de un punto, se recurre a un tercer lector (ETS, 2007). Las puntuaciones se promedian y se informan en aumentos de medio punto, es decir, 4.0, 4.5, 5.0; también se proporcionan los percentiles. Los cambios de percentil por nivel de puntuación son, en verdad, asombrosos en la mitad de la escala, pues aumentan de 15 a 20 puntos por cada aumento de medio punto en el nivel de la puntuación. Por ejemplo, las puntuaciones de 3.0, 3.5 y 4.0 corresponden a los percentiles 11, 30 y 49, respectivamente. El GRE ha manifestado su intención de introducir la calificación automatizada (véase 23a») de los ensayos en el futuro (ETS, 2012a).

¡Inténtalo!

Para ver reactivos muestra del GRE Prueba General, entra a http://www.ets.org/gre/revised_general/prepare/

Escala de puntuaciones y normas del GRE

Desde sus orígenes y hasta agosto de 2011, el GRE Prueba General usó la conocida escala de puntuaciones 200–800 con $M = 500$ y $DE = 100$, la misma que el SAT. A partir de agosto de 2011, el GRE cambió a una escala que va de 130 a 170; al menos hasta ahora, las publicaciones del GRE no han destacado una media y una desviación estándar para la nueva escala, pero la inspección de los cuadros de normas revela que la escala se aproxima a una con $M = 150$ y $DE = 10$. Una nota más bien oscura indica que la *DE* se fijó en el extraño número de 8.75 (ETS, 2012a, p. 18). Es interesante que Law School Admission Test [Prueba de Admisión a la Escuela de Derecho] haya usado una escala con $M = 150$ y $DE = 10$ desde 1991 (excepto porque LSAT advierte que sus escalas van de 120 a 180), aunque también haya usado la escala 200-800 en sus primeros días con un breve coqueteo con una escala 0-50 (Dalessandro, Stilwell, Lawlor, & Reese, 2010). Las publicaciones del GRE ofrecen cuadros de concordancia para convertir puntuaciones entre las antiguas escalas 200-800 y las nuevas 130-170.

¡Inténtalo!

Escribe "GRE concordance" en cualquier buscador de internet para obtener una copia de los cuadros. ¿A qué corresponde una puntuación de 150 de la nueva escala en la escala antigua en Verbal y Cuantitativo? ¿Qué harías con una diferencia tan alarmante?

¿Por qué no todas estas personas se reúnen y establecen una sola escala de puntuaciones para usarla, digamos, durante los próximos 100 años? Nos están volviendo locos, al menos durante esta fase de transición, y estamos gastando montones de papel imprimiendo toda clase de cuadros de concordancia. He aquí un par de sugerencias. Si no te gusta la escala 200-800, sólo quita el dígito de las unidades (que siempre es cero) y listo, tienes la escala de puntuaciones T con *M* = 50 y *DE* = 10 (véase figuras 3-10a y 3-10b). Si tomamos la escala 120-180 y restamos 100 de todas las puntuaciones, tenemos otra vez una escala de puntuaciones T, que es casi el "estándar de la industria" del campo de las pruebas de personalidad. Otra alternativa sería que cada una usara la muy conocida escala de CI, pero es improbable que eso ocurra porque la gente no quiere dar a entender que estas pruebas miden IQ, aunque en realidad hacen esto.

Las normas de rangos percentiles del GRE constituyen un clásico ejemplo de las "normas del usuario". Regresa a la página 67a» para encontrar la descripción de este tipo de normas, que se basan sólo en las personas que, en realidad, respondieron la prueba en cierto periodo de tiempo. De modo que las normas del GRE se actualizan cada año con base en los examinados de los tres años más recientes (sólo un año en el caso de la prueba y las escalas recientemente revisadas). Cada año surge un nuevo conjunto de rangos percentiles, en el que se agregan nuevos casos del año anterior inmediato y se excluyen casos de hace tres años.

Confiabilidad y validez «250a

Un número en verdad asombroso de estudios ha examinado la confiabilidad y validez del GRE Prueba General. Es fácil tener dolor de cabeza al revisar cuadro tras cuadro de coeficientes de confiabilidad y validez buscando diferencias menores entre una prueba y otra o entre un grupo y otro. Sin embargo, en general, emerge una imagen muy clara de todos estos estudios y reseñas; primero identificaremos las fuentes clave de datos y después proporcionaremos un conjunto de generalizaciones.

Quizá la revisión clásica de la investigación sobre el GRE es la de Kuncel, Hezlett y Ones (2001), quienes realizaron un metaanálisis que incluyó 1753 muestras y más de 80 000 alumnos de posgrado. Incluso el material de la editorial emplea esta fuente primaria de la validez del GRE (ETS, 2009). El metaanálisis de 2001 fue complementado con otro en 2010 basado en casi 100 estudios y 10 000 estudiantes (Kuncel, Wee, Serafin, & Hezlett, 2010). Burton y Wang (2005) presentaron una revisión interna muy completa. Bridgeman, Burton y Cline (2008) usan, en esencia, la misma base de datos, pero la analizan de una manera un poco diferente. Cada guía anual del GRE (p. ej., ETS, 2012a) ofrece información sobre la confiabilidad y validez, pero la de 2005-06 (ETS, 2005) presentó un resumen particularmente bueno. Los datos de la prueba Escritura analítica provienen, en su mayor parte, de Schaeffer, Briel y Fowles (2001) y ETS (2009).

Antes de presentar algunas generalizaciones, una nota de precaución. Muchos de los resúmenes existentes dan información de la "antigua" prueba Analítica, que ahora es irrelevante, mientras que hay poca información disponible sobre la prueba Escritura

analítica. Algunas fuentes sólo hablan de las pruebas Verbal y Cuantitativa; hay aun menos información acerca de la Prueba General "revisada". Por ello, vamos a suponer que la información de las "antiguas" pruebas Verbal y Cuantitativa se puede generalizar a las versiones revisadas de estas pruebas, pero no suponemos lo mismo en lo que respecta a la "antigua" versión de la prueba Analítica en relación con Escritura analítica.

Las pruebas Verbal y Cuantitativa producen puntuaciones de alta confiabilidad. Los coeficientes de confiabilidad derivados mediante distintos métodos, por ejemplo, KR-20 y procedimientos de TRR, por lo general se ubican dentro del rango de .90 a .93. Estos datos son consistentes con los de la confiabilidad de una gran cantidad de medidas de capacidad mental general que se califican de manera objetiva. Hasta aquí, no hay ninguna sorpresa. Los errores estándar de medición (EEM) son de 2 a 3 puntos en promedio. Aplicando la metodología de TRR (véase p. 97a»), se pueden informar los EEM de varias partes del rango de puntuaciones. Los EEM tienden a ser menores en los rangos superiores y mayores en los rangos inferiores.

Analizar la confiabilidad de la prueba Escritura analítica presenta algunos problemas particulares, porque requiere del juicio humano (confiabilidad interjueces), y sólo hay dos ejercicios de escritura. La investigación con otras pruebas muestra que la confiabilidad depende más del número de tareas que del número de jueces (Hayes, Hatch, & Silk, 2000; Lee & Kantor, 2007). Así, considerar sólo la confiabilidad interjueces de Escritura analítica exagera su confiabilidad operacional. De hecho, esta confiabilidad es respetable, por lo general, si es mayor de .75 (lo cual está muy por debajo del .95 de las pruebas Verbal y Cuantitativa). Pero la confiabilidad es menor de .75 cuando las dos tareas y jueces se toman en cuenta, lo cual no es muy bueno.

Al igual que todas las pruebas, la pregunta más importante para el GRE Prueba General tiene que ver con su validez. El primer paso para resolver esta pregunta implica determinar un criterio. ¿Qué define el éxito en los programas de posgrado? ¿Las notas en el programa? ¿O las valoraciones de los profesores del éxito general independientemente de las notas? ¿O la productividad final como experto medida por el número de publicaciones 10 años después de concluir el programa? ¿O quizá algún otro criterio?

El criterio más común empleado para los estudios de validez del GRE es el GPA de primer año en el programa de posgrado (FYGPA). Literalmente miles de estudios han usado este criterio. Además, un número menor pero aún considerable de estudios ha empleado otros criterios, incluyendo el GPA de fin del programa y valoraciones generales de los profesores. El cuadro 9-11 muestra un pequeño resumen de información clave sobre la validez.

Cuadro 9-11. Generalizaciones acerca de la validez del GRE Prueba General

	Correlación con el FYGPA	
	Observada	**Corregida**[a]
Cualquier prueba sola (V, Q, A)	.25-.30	.35-.45
R múltiple con 2-3 pruebas	.35	.45
GPA de estudiantes universitarios (GPAEU)	.35	.45

R múltiple con 2-3 pruebas + GPAEU	.45	.55

[a] Corregida de acuerdo con la restricción de rango y confiabilidad imperfecta.

¡Inténtalo!

¿Se te ocurre algún otro criterio que pueda usarse para definir el éxito en programas de posgrado distinto de los que acabamos de mencionar?

El resumen general de toda esta información sugiere las siguientes generalizaciones:

1. Las tres pruebas del GRE Prueba General tienen un modesto grado de validez predictiva.

2. Las tres pruebas (V, Q, A) tienen una validez predictiva notablemente parecida.

3. Los compuestos, V + Q o V + Q + A, son ligeramente mejores, pero sólo ligeramente, como predictores que cualquiera de estas pruebas sola.

4. El GPA de estudiantes universitarios (GPAEU) es un mejor predictor que cualquiera de las pruebas y más o menos igual que el compuesto (VQA) en poder predictivo.

5. La combinación de las tres pruebas *y* el GPA de estudiantes universitarios (VQAEU) es claramente superior a cualquier prueba sola o el GPAEU solo en la predicción del FYGPA. La validez incremental es cercana a .10.

6. Las últimas generalizaciones son aproximadamente las mismas en todos los tipos de departamentos.

7. No estamos examinando el GRE Pruebas por Temas; sin embargo, debemos señalar que los datos muestran que estas pruebas son mejores predictores del FYGPA que las pruebas Generales, en promedio, de los distintos campos por cerca de .10. Además, en algunos campos, las pruebas por Tema son mejores predictores que el GPA de estudiantes universitarios, mientras que en otros, los dos predictores son aproximadamente iguales.

8. Se requiere más información acerca de la validez predictiva de la prueba Escritura analítica. Algunos datos sugieren que funciona casi tan bien como la antigua prueba Analítica (ETS, 2007), pero no tenemos tanta información sobre ello como del resto de las pruebas.

Aquí hay algunos puntos para ayudar a tener una perspectiva de toda esta información. Primero, la gente que está a favor del uso de las pruebas destaca que estas correlaciones son muy significativas y que cualquier grado de validez es mejor que la selección al azar, mientras que la gente que rechaza las pruebas destaca que las pruebas aún dejan mucha varianza sin explicar en el FYGPA. Segundo, por un lado, es notable que una prueba de sólo 60 min (cualquiera de las pruebas Generales del GRE) hace un trabajo casi tan bueno al predecir el FYGPA como el resumen de un trabajo entero de cuatro o cinco años, es decir, el UGPA. Por otro lado, ninguno de estos predictores, incluyendo el UGPA, es un gran predictor del FYGPA.

Pruebas de selección en el Ejército y los negocios

El Ejército y los negocios constituyen la tercera aplicación importante de las pruebas grupales de capacidad mental. Los fundamentos son los mismos que los de las dos primeras aplicaciones. La intención es elegir individuos con una capacidad mental general suficiente para desempeñarse con éxito en un trabajo o en un programa de entrenamiento. El cuadro 9-12 enumera algunas de las pruebas más usadas de esta categoría. Un ejemplo es el *General Aptitude Test Battery* [Batería de Pruebas de Aptitud General] (GATB), financiado por el Departamento del Trabajo de EUA. Los negocios usan una gran variedad de pruebas. El *Differential Aptitude Test* [Prueba de Aptitudes Diferenciales] (DAT) se usa principalmente en bachillerato, pero también es importante en la industria. Algunas pruebas de esta categoría derivaron de las pruebas Otis, por ejemplo, el *Wonderlic Personnel Test* [Prueba de Personal Wonderlic]. Los empresarios a veces usan pruebas de capacidades o conocimientos específicos (véase, p. ej., las pruebas NOCTI en el capítulo 11) relacionados con un puesto particular. Sin embargo, muchos empresarios y el Ejército de EUA usan pruebas de capacidad mental general, como las que revisamos en este capítulo. Una extensa investigación ha documentado la validez de estas pruebas para seleccionar empleados en una gran variedad de puestos (Kuncel & Hezlett, 2010; Schmidt, Ones, & Hunter, 1992). A continuación examinamos la prueba principal que se usa en escenarios militares, seguida de una muy empleada en los negocios.

Cuadro 9-12. Algunas pruebas grupales de capacidad mental muy usadas en los negocios y el Ejército

Iniciales	Nombre completo	Fuente
GATB[a]	General Aptitude Test Battery	Departamento del Trabajo de EUA
DAT	Differential Aptitude Test	Pearson Assessments
WPT	Wonderlic Personnel Test	Wonderlic, Inc.
ASVAB	Armed Services Vocational Aptitude Battery	Departamento de Defensa de EUA

[a] Algunas formas recientes del GATB se conocen como *Ability Profiles*, disponible por medio de O*NET, proyecto que se describe en la página 394a».

Batería de Aptitud Vocacional de las Fuerzas Armadas

El **Armed Services Vocational Aptitude Battery** [Batería de Aptitud Vocacional de las Fuerzas Armadas] (**ASVAB**) se usa en todas las áreas militares de EUA para la selección de reclutas y su ubicación en puestos específicos. El Ejército aplica cerca de dos millones de ASVAB al año. Las raíces de esta prueba datan de la Primera Guerra Mundial; recordemos, de la historia de las pruebas que vimos en el capítulo 1, que el uso inicial de

pruebas de capacidad mental de aplicación grupal, el Army Alpha y Army Beta, fue examinar a los reclutas del Ejército. Después, todas las áreas del Ejército empezaron a usar estas pruebas de manera cotidiana; durante muchos años, cada área del servicio empleó su propia prueba, de las cuales quizá la más famosa fue *Army General Classification Test* [Prueba de Clasificación General del Ejército] (AGCT). A inicios de 1976, todos los servicios empezaron a usar el único ASVAB. En Larson (1994) y Sands y Waters (1997) se puede encontrar una breve historia del ASVAB y sus predecesores. A finales de la década de 1990, un equipo emprendió el proyecto de crear una versión adaptada para computadora del ASVAB. Sands, Waters y McBride (1997) presentan una descripción particularmente completa de este proyecto. En la actualidad, la prueba está disponible en ambas formas: adaptada para computadora y lápiz y papel.

Estructura y reactivos

El cuadro 9-13 bosqueja la estructura actual del ASVAB. Cuenta con ocho subpruebas con un tiempo total de aplicación de aproximadamente 3 hrs. Las ediciones previas tenían otras subpruebas además de las de la edición actual, por ejemplo, Velocidad de codificación, Ensamble de objetos y Operaciones numéricas.

Cuadro 9-13. Estructura del ASVAB

Subprueba		**Reactivos**	**Tiempo (min)**	**Descripción**
RA	Razonamiento aritmético	30	36	Problemas aritméticos con palabras
CP	Conocimiento de palabras	35	11	Sinónimos, significado de palabras en contexto
CM	Conocimiento matemático	25	24	Matemáticas de nivel bachillerato
CP	Comprensión de párrafos	15	13	Lectura de párrafos
CG	Ciencia general	25	11	Ciencia: física (13 reactivos), biología (12 reactivos)
CM	Comprensión mecánica	25	19	Principios físicos y mecánicos
IE	Información de electrónica	20	9	Principios de electrónica, radio, electricidad
AC	Información de autos y compras	25	11	Términos y prácticas relacionadas con autos, herramientas, compras

Fuente: Adaptado de Larson (1994) y de Sands y Waters (1997).

Es instructivo examinar la lista de subpruebas del ASVAB. Primero, encontramos que es una mezcla de contenido muy extraña, en especial tratándose de una prueba que se denomina de aptitud. Algunas de las subpruebas, por ejemplo, Conocimiento de palabras y Razonamiento aritmético, son entradas clásicas de las pruebas de aptitud general, mientras que otras, por ejemplo, Información electrónica e Información de autos y

compras, son áreas de conocimiento muy específicas. ¿Por qué se encuentran éstas en una prueba de "aptitud"? Segundo, un punto relacionado: nos preguntamos acerca de cuáles son los rasgos que mide el ASVAB. ¿En realidad son ocho capacidades diferentes abordadas por la prueba? No, la investigación indica que sólo hay unas pocas capacidades subyacentes (ASVAB Career Exploration Program, 2010; Sands & Waters, 1997). Este resultado, sin duda, es consistente con otras investigaciones sobre la estructura de las capacidades humanas. De hecho, algunos informes recientes sobre el ASVAB proporcionan puntuaciones de los compuestos Verbal, Matemáticas y Ciencia/Técnica. Tercero, podemos notar que algunas subpruebas son muy cortas, de las cuales esperaríamos que tuvieran una confiabilidad bastante limitada; en efecto, así es.

¡Inténtalo!

Para ver ejemplos de reactivos de cada subprueba del ASVAB, entra en http://www.official-asvab.com/samples_coun.htm

Características técnicas

Aunque el ASVAB produce puntuaciones de cada subprueba que aparece en el cuadro 9-13, tal vez la más importante es la de **AFQT**. Antes de que se adoptara el ASVAB, existía el **Armed Forces Qualification Test** [Prueba de Calificación de las Fuerzas Armadas], que constaba de reactivos verbales, cuantitativos y espaciales, pero se abandonó en favor del ASVAB; sin embargo, se conservaron las iniciales AFQT para designar una puntuación especial, que consiste en una combinación: 2(CP + CP) + RA + CM. De ahí que sea una medida tipo "*g*" que combina los dominios verbales y cuantitativos. La puntuación AFQT se usa, en realidad, para tomar decisiones de selección de reclutas en el Ejército. La puntuación se convierte en una lista de categorías que aparece en el cuadro 9-14.

Cuadro 9-14. Categorías de las puntuaciones AFQT obtenidas del ASVAB

Categoría	Percentiles	Categoría	Percentiles
I	93-99	IVA	21-30
II	65-92	IVB	16-20
IIIA	50-64	IVC	10-15
IIIB	31-49	V	0-9

La denominación de las categorías del AFQT parece un poco extraña, pero tiene una calidad casi mística dentro de los círculos militares. De manera general, la categoría V descalifica a una persona de la selección en el Ejército, y la categoría IV tiene una selección muy restringida. «254a

Podríamos esperar que las normas de percentiles del ASVAB fueran normas del

usuario, como las del SAT, ACT y GRE; sin embargo, éste no es el caso. Las normas del ASVAB provienen de un estudio nacional controlado de manera cuidadosa de 6000 hombres y mujeres de edades entre 18 y 23 años. Se trató del National Longitudinal Study of Youth [Estudio Nacional Longitudinal de la Juventud] (NLSY97) realizado en 1997 (Defense Manpower Data Center, 2004). Por ello, el ASVAB tiene normas que deben ser razonablemente representativas de la población nacional, al menos la de 1997.

En las versiones adaptada para y de lápiz y papel, las confiabilidades de las subpruebas del ASVAB son, en gran parte, predecibles a partir de su extensión (véase Moreno & Segall, 1997). Como regla general, las pruebas cortas no son muy confiables; entre las subpruebas del ASVAB, Comprensión de párrafos (15 reactivos) es especialmente problemática por su confiabilidad estimada a menudo por debajo de .70. El resto de las subpruebas tiene confiabilidades alrededor de .80. Desde luego, el compuesto AFQT es muy confiable, con coeficientes arriba de .90 en la mayoría de estudios.

Los escenarios militares ofrecen muchas oportunidades de estudiar la validez del ASVAB como predictor de éxito en diversos programas de entrenamiento. Los estudios en numerosos programas de entrenamiento militar demuestran una validez respetable, en especial de la puntuación compuesta AFQT. En muchos programas de entrenamiento (p. ej., controlador del tráfico aéreo, técnico en electrónica, encargado de la radio), el compuesto AFQT mostró correlaciones corregidas por rango con un promedio de .60 con las notas finales de la escuela de entrenamiento y correlaciones un poco más bajas, pero aún respetables, con el desempeño en el trabajo (Welsh, Kucinkas, & Curran, 1990; Wolfe, Moreno, & Segall, 1997). A menudo, una sola prueba, como Conocimiento de palabras o Razonamiento Aritmético, no está muy atrás del compuesto AFQT en validez predictiva. Prácticamente todos los que han reseñado el ASVAB (véase, p. ej., Murphy, 1984) han señalado con disgusto la falta de *validez diferencial* de las distintas subpruebas.

Prueba de Selección de Personal Wonderlic

En notable contraste con el AFQT de puntuaciones múltiples, se encuentra la *Prueba de Personal Wonderlic* [Wonderlic Personnel Test] (WPT; Wonderlic, 2002, 2007),[3] quizá la prueba de capacidad mental más usada en el contexto de recursos humanos, es decir, para la selección de empleados. El WPT es la esencia misma de la sencillez: cada forma consta de 50 reactivos aplicados con un límite de tiempo de sólo 12 min y produce una sola puntuación general. Los materiales de WPT descaradamente pretenden medir la inteligencia general, es decir, "*g*" (véase la discusión de "*g*" en pp. 178-179a»).

El WPT apareció por primera vez en 1938, cuando las pruebas de inteligencia como el Stanford-Binet y el Otis empleaban una sola puntuación; de hecho, en apariencia y estructura general actual, el WPT es prácticamente indistinguible de las primeras versiones de las pruebas Otis. Mientras que la mayoría de las pruebas de capacidad mental adoptaron después un método de puntuaciones múltiples, el WPT mantuvo con firmeza su método de puntuación única. El WPT incluye una mezcla (o una mezcolanza)

de reactivos de los dominios verbales, cuantitativos y espaciales. Véase los ejemplos de reactivos de opción múltiple en el cuadro 9-1.

Debemos tener en mente que Wonderlic, Inc. publica varias pruebas además del WPT, incluyendo medidas del campo de la personalidad. Algunas de estas pruebas incorporan el nombre de Wonderlic, por lo que, cuando alguien dice "usamos el Wonderlic", es mejor indagar con exactitud a qué se refiere. Nuestra descripción se enfoca sólo en el WPT.

El WPT cuenta con una confiabilidad alta. Las confiabilidades de consistencia interna, por lo general, superan el .90, aunque estas cifras están infladas un poco por la velocidad que caracteriza a la prueba. Sin embargo, las confiabilidades de test-retest también son altas, lo cual es destacable para una prueba de 12 min. La información de la validez se concentra en dos tipos de estudios: correlaciones con otras pruebas de capacidad mental, sobre todo con las que son más extensas, y correlaciones con diversos índices de desempeño en el trabajo. Los estudios de este último tipo, desde luego, son cruciales para los usos del WPT en recursos humanos. Las correlaciones con otras pruebas de capacidad mental, incluyendo las de aplicación individual como el WAIS, tienden a ser bastante altas, a menudo muy cerca del nivel de confiabilidad de estas pruebas. Esto, otra vez, es un resultado notable para una prueba de 12 min en comparación con pruebas que requieren 1 hr para su aplicación.

Las correlaciones del WPT con índices de desempeño en el trabajo, por ejemplo, valoraciones de los empleados hechas por supervisores, tienden a estar alrededor de .35. Cuando estas correlaciones se corrigen por confiabilidades imperfectas en las medidas (véase estos procedimientos en pp. 117-118b») y por restricción en el rango (véase estos procedimientos en pp. 81-83a»), prácticas comunes en la literatura de investigación, se ubican alrededor de .55. ¿Qué debemos pensar de estas correlaciones? Vale la pena señalar dos puntos. Primero, las correlaciones con las pruebas tipo "*g*", como el WPT, tienden a ser mejores predictores del desempeño en el trabajo que cualquier otra (a excepción del desempeño previo en el trabajo). Una enorme literatura de investigación apoya este punto (véase, p. ej., Kuncel & Hezlett, 2010; Schmidt & Hunter, 2004). Segundo, aunque hacer las correcciones que mencionamos es útil para propósitos de análisis psicométricos, el hecho es que, en los usos actuales en situaciones prácticas, siempre lidiamos con confiabilidades imperfectas y rangos restringidos. Con esto en mente, mientras una prueba como el WPT puede tratar de venderse como una alternativa mejor que distintos predictores del desempeño en el trabajo, una correlación de .30 (entre el desempeño en la prueba y en el trabajo) no es muy alta.

¡Inténtalo!

Un detalle curioso acerca del WPT es que todos los jugadores de la National Football League (NFL) que estaban en venta lo respondieron. Los resúmenes de estos resultados por posición (y en algunos casos por jugador) están disponibles. Introduce "Wonderlic y NFL" en cualquier buscador de internet para encontrar qué posición tiene la puntuación promedio más alta en el Wonderlic.

He aquí una pregunta: ¿por qué la gente insiste en usar pruebas de 3 hrs cuando pruebas mucho más cortas, digamos, de 30 min, parecen hacer predicciones igual de buenas del éxito en la escuela y el trabajo? Incluso el Wonderlic de 12 min parece ser tan bueno como el ASVAB de 3 hrs. Cualquiera de las subpruebas de 30 min del ASVAB parece ser tan buena como la batería completa. La prueba verbal del SAT (Lectura crítica) parece ser tan buena como el complejo SAT entero.

Pruebas de capacidad mental culturalmente neutrales

Éste es uno de los grandes dilemas del mundo de la psicometría. Por un lado, las definiciones comunes de inteligencia hacen hincapié en la capacidad para tener éxito en el medio, lo cual sugiere evaluar mediante símbolos, convenciones y artefactos ordinarios de la cultura. Por otro lado, desearíamos una medida "pura" de inteligencia que no estuviera limitada por las particularidades de una cultura. Este deseo ha adquirido el carácter de urgente conforme ha aumentado la preocupación por el desempeño de los grupos minoritarios. Sin embargo, debemos señalar que la búsqueda de la medida "pura" por parte de los psicólogos precede a esta preocupación contemporánea, pues se originó, como veremos, en la segunda mitad de la década de 1930, mientras que la preocupación por el desempeño de grupos minoritarios surgió en la de 1960.

Pruebas como el Stanford-Binet, las escalas Wechsler, las series Otis y el SAT están, obviamente, impregnadas de las prácticas culturales de Occidente; en gran parte, son verbales, específicamente en inglés estándar. ¿Existen modos de medir la inteligencia sin depender de una lengua y cultura particulares? Ésta es la pregunta que atañe a lo que llamamos prueba libre de cultura o **prueba culturalmente neutral**; este término es el que se prefiere en la actualidad, aunque algunas fuentes emplean el término reducido en cultura. Libre de cultura fue la denominación original de estas pruebas, pero no tuvo que pasar mucho tiempo para que fuera evidente que ninguna prueba podría estar por completo libre de todo atributo cultural. Por ejemplo, usar papel, suponer una orientación de izquierda a derecha y de arriba a abajo en la página, dar respuestas directas a las preguntas y aceptar restricciones de tiempo son prácticas ligadas a la cultura. Sin embargo, quizá podemos crear una prueba que sea neutral y equitativa en términos culturales. Consideremos algunos esfuerzos para conseguirlo. Debemos señalar que, aunque tratamos este tema en relación con las pruebas grupales de capacidad mental, también es pertinente para las de aplicación individual.

Matrices Progresivas de Raven

Tal vez el ejemplo más conocido de una prueba que pretende ser culturalmente neutral es **Matrices Progresivas de Raven** (MPR). Se trata de un ejemplo importante porque se cita con mucha frecuencia en la literatura de investigación sobre la inteligencia. Recordemos de nuestra discusión sobre las teorías de la inteligencia el papel central de la inteligencia general, "*g*"; muchas personas consideran el MPR como una medida de "*g*" de especial importancia. A menudo es un punto de referencia de los estudios analítico-factoriales de la inteligencia; por lo tanto, los estudiantes de las pruebas psicológicas deben tener un conocimiento básico de esta prueba.

Muchas fuentes se refieren a esta prueba como "el Raven" o "las Matrices de Raven"

como si se tratara de una sola prueba. Sin embargo, las matrices de Raven constituyen en realidad tres series diferentes de pruebas, como se resumen en el cuadro 9-15. Primero, las *Matrices Progresivas en Color* (MPC), diseñadas para niños pequeños y, en general, para la parte inferior de la distribución de la inteligencia; la prueba usa colores para aumentar el nivel de interés. Segundo, las *Matrices Progresivas Estándar* (MPE); se trata de la versión clásica que consta de 60 reactivos. Está dirigida a personas de la zona media del espectro de capacidad mental. Su edición más reciente es la versión "Extendida Plus", publicada en 1998. Tercero, las *Matrices Progresivas Avanzadas* (MPA), diseñadas para el 20% superior de la distribución de capacidad mental. Una sola prueba, *Matrices Progresivas de Raven*, precursora del MPE, se publicó en 1938. Recordemos del capítulo 7 que, en 1940, Cattell publicó su artículo seminal, en el que propuso una prueba libre de cultura basada en su mayor parte en reactivos tipo matriz.

Cuadro 9-15. Las tres versiones de Matrices Progresivas de Raven

Título	**Tiempo (min)**	**Grupo meta**
Matrices Progresivas en Color (MPC)	15-30	5-11 años, 20% inferior
Matrices Progresivas Estándar (MPE) (Versión Extendida Plus)	20-45	6-16 años, población general
Matrices Progresivas Avanzadas (MPA)	40-60	12-17+ años, 20% superior

La figura 9-7 muestra reactivos ficticios de Raven en distintos niveles de dificultad. El "tronco" del reactivo muestra un patrón (matriz) con una parte faltante, y la tarea del examinado es elegir la opción que completa el patrón. Lo importante de nuestra presentación aquí es observar la naturaleza de los **reactivos tipo matriz**.

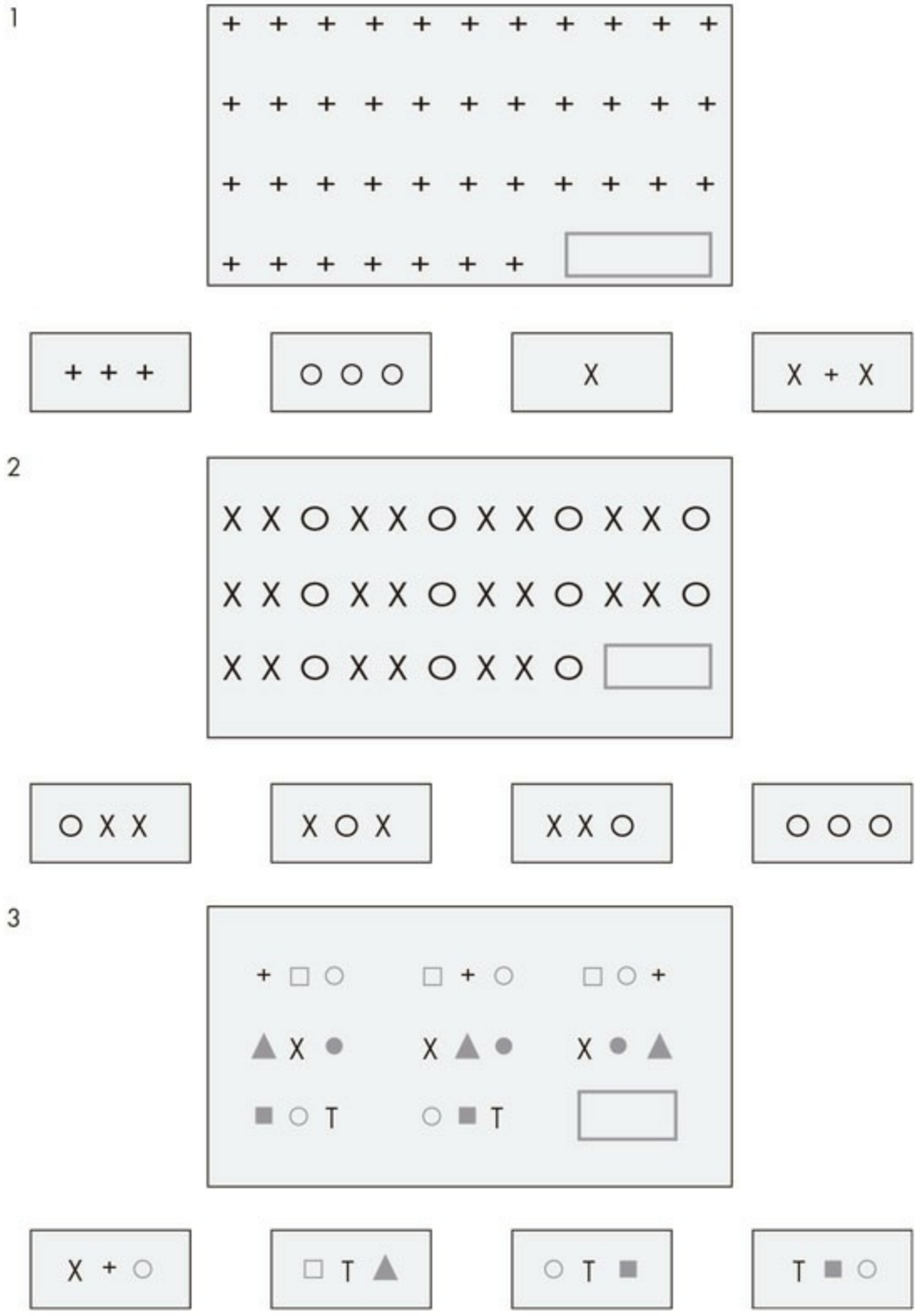


Figura 9-7. Reactivos ficticios tipo Raven.

El Raven tiene varias características deseables. Es por completo no verbal; incluso las

instrucciones se pueden dar con pantomima si es necesario. De ahí que sea atractivo usarlo con individuos de diferentes culturas y lenguas, o con dificultades físicas (excepto si la dificultad le impide ver). Es de opción múltiple, por lo que es fácil calificarlo. Su aplicación no requiere un entrenamiento especial y puede usarse con individuos o grupos. Además, es razonablemente breve.

¿Por qué no se usa más el Raven en escenarios prácticos? La respuesta, quizá, radica en su mezcolanza de materiales confusos y descoordinados. Hay tres niveles, cada uno con un título diferente. Existen, al menos, cinco manuales publicados por separado acompañados de diversos suplementos. Hay una gran variedad de grupos de estandarización. Mucho más importante es la evidencia en conflicto relacionada con el rasgo o rasgos que mide la prueba. El manual de Raven hace hincapié en la medición de la "deducción de relaciones" de la "*g*" de Spearman (Raven, Court, & Raven, 1993), con lo cual concuerdan varios críticos (p. ej., Llabre, 1984; Vernon, 1984). Por otro lado, algunos autores (p. ej., Esquivel, 1984; Gregory, 2011) señalan que los estudios analítico-factoriales identifican varios rasgos diferentes, incluso en un solo nivel. Podemos suponer cuáles son estos rasgos a partir de los reactivos ficticios de la figura 9-7; el primer reactivo parece basarse principalmente en la percepción, mientras que el tercero demanda razonamiento analógico. Entre los hallazgos se encuentra una especie de capacidad figurativa/espacial. Por último, el manual de Raven hace referencia continua a las pruebas guía de vocabulario; sin embargo, dentro de lo complejo del Raven, los dos tipos de pruebas no están bien coordinadas. A pesar de ello, el Raven, en sus varias formas, es un instrumento muy citado en el campo de las pruebas psicológicas.

Otras pruebas culturalmente neutrales y algunas conclusiones

De ningún modo el Raven es el único intento de ofrecer una prueba culturalmente neutral, sino que hay numerosos ejemplos. Antes, en la descripción de las teorías de la inteligencia, mencionamos el artículo de 1940 de Cattell en el que anunció una prueba culturalmente neutral. La prueba consistió en su mayor parte de reactivos tipo matriz. Con el tiempo, Cattell elaboró la prueba en la publicación regular *Culture Fair Intelligence Test* [Prueba de Inteligencia Culturalmente Neutral] (CFIT). El trabajo inicial con estos reactivos tipo matriz fue parte del fundamento de Cattell para distinguir entre inteligencia fluida y cristalizada. Se suponía que el CFIT abordaría la dimensión fluida, mientras que pruebas más orientadas hacia lo verbal se encargarían de la dimensión cristalizada. Muchos otros autores han construido pruebas empleando reactivos tipo matriz, relaciones figurativas, diseños geométricos y otros estímulos no verbales. La figura 9-8 presenta ejemplos de estos reactivos. La esperanza, casi siempre, es que la prueba sea culturalmente neutral.

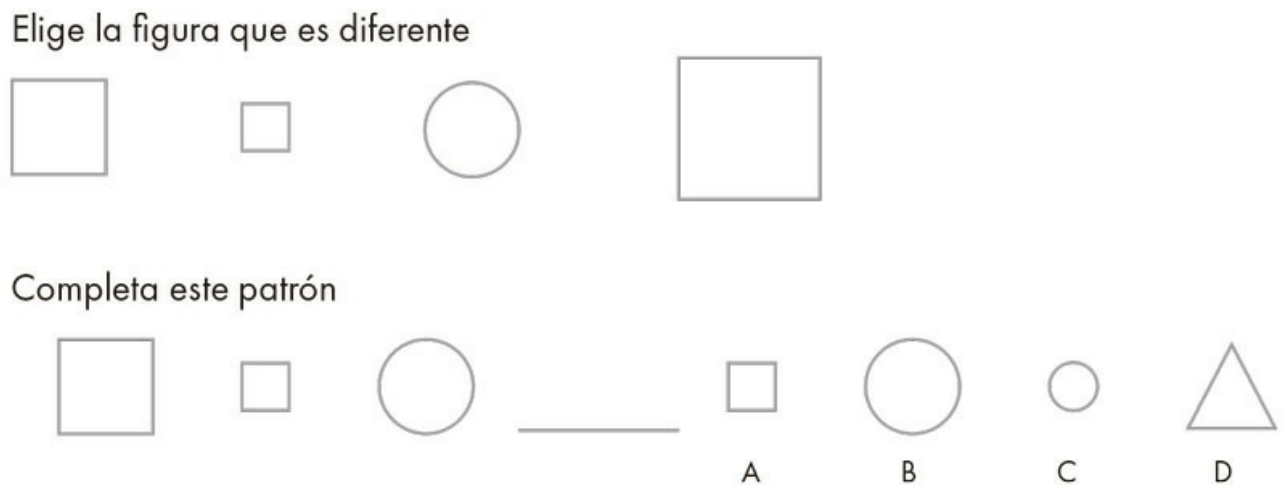


Figura 9-8. Ejemplos de reactivos no verbales y otros distintos de los de tipo matriz.

Debemos señalar que las tareas cognitivas elementales (TCE) usadas en los modelos de procesamiento de información de la inteligencia (véase pp. 185-187a») se han propuesto como medidas culturalmente neutrales (incluso libres de cultura) de la inteligencia, lo cual es una noción intrigante.

Sin embargo, la aplicación de las TCE ha estado confinada por largo tiempo a los usos de laboratorio; además, tienen aún mucho que mostrar acerca de su valor como medidas de inteligencia.

Las siguientes tres conclusiones surgen de la revisión de un trabajo sobre pruebas culturalmente neutrales. Primero, las pruebas tienden a ser medidas de capacidad de razonamiento figurativo y espacial; esto puede ser de alguna utilidad, sobre todo cuando no hay una posibilidad razonable de usar una medida más convencional de capacidad mental. Sin embargo, estas pruebas, evidentemente, no miden las mismas capacidades que las pruebas de referencia del funcionamiento intelectual general, de la manera tradicional como lo hacen las de Wechsler y Otis. Debemos ser cautelosos ante las afirmaciones de estos autores y editoriales acerca de "una prueba no verbal, culturalmente neutral de inteligencia" por dos cuestiones. Primero, "la inteligencia" a la que se hace referencia en la afirmación es de una variedad mucho más circunscrita; si la afirmación implica que esta prueba no verbal puede ser sustituto de, digamos, el WISC, se trata de una afirmación engañosa. Segundo, podemos determinar por la simple inspección que una prueba es, en gran parte, no verbal, pero no podemos determinar del mismo modo que es culturalmente neutral. El hecho de que una prueba sea no verbal no la hace de manera automática culturalmente neutral; para saber si una prueba lo es se requiere investigación que muestre que funciona de modo equivalente en varias culturas.

Segundo, cuando se usan para predecir el éxito escolar o laboral, las pruebas que son primordialmente medidas de razonamiento figurativo o espacial son claramente inferiores en relación con las pruebas de orientación verbal. Además, las pruebas figurativas/espaciales agregan poco, si lo hacen, poder predictivo a las pruebas verbales

sencillas cuyo fin es precisamente la predicción. (Hay algunas excepciones muy limitadas de esta generalización, p. ej., predecir el éxito en arquitectura.) La razón de la superioridad de las medidas verbales quizá es muy sencilla. Las actividades académicas y laborales tienen más demandas verbales que figurativas/espaciales.

Tercero, no se ha cumplido la esperanza de que las pruebas culturalmente neutrales eliminen las diferencias en las puntuaciones promedio entre grupos mayoritarios y minoritarios o entre grupos de distintas culturas. Algunas investigaciones muestran que estas diferencias se reducen un poco en las pruebas no verbales, en comparación con las pruebas más convencionales con mucha carga verbal. Otras investigaciones muestran que las diferencias grupales son aproximadamente las mismas en pruebas con fuerte carga verbal y las no verbales. La búsqueda del vellocino de oro habrá de continuar.

Pruebas de inteligencia para microculturas

En el otro extremo de los intentos por construir pruebas de inteligencia culturalmente neutrales están las que se basan en subculturas sumamente específicas, lo que podemos llamar microcultura. Con frecuencia escuchamos acerca de tales pruebas de "inteligencia"; por ejemplo, una de ellas podría desarrollarse con terminología de navegación (puerto, virada, estribor, etc.), de beisbol (toque de bola, fly de sacrificio, corredor emergente, etc.) o del sistema de transporte subterráneo (metro, torniquete, andén, etc.). A menudo, estas pruebas se presentan para desacreditar las pruebas convencionales de capacidad mental, como el WAIS o el SAT. Algún bromista señalará que una persona con un CI de 150 en el WAIS (que vive en Chicago) falló en una prueba sobre el metro de la Ciudad de Nueva York, mientras que otra con un CI de 90 (que vive en el Bronx) pasó la prueba. Esto implica que el CI del WAIS no es importante, y los medios están encantados de pregonar estos informes; sin embargo, es claro que esta implicación es una tontería. Sabemos mucho acerca de lo generalizable que puede ser el CI del WAIS, así que debemos hacer las mismas preguntas acerca de la prueba del metro que acerca del WAIS: ¿A qué otra conducta se puede generalizar el desempeño en la prueba del metro? ¿Con qué se correlaciona? ¿Cuáles son las normas de la prueba? ¿La puntuación es confiable? Por lo general, esta información no está disponible en estas pruebas para microculturas.

Generalizaciones acerca de las pruebas grupales de capacidad mental

Nuestro examen de las pruebas grupales de capacidad mental sugiere las siguientes seis generalizaciones, las cuales deben matizarse con información específica de cada prueba y sus aplicaciones particulares. Sin embargo, algunas tendencias son claras en varias de las pruebas que hemos revisado.

Resumen de puntos clave 9-3

Generalizaciones acerca de las pruebas grupales de capacidad mental

- Contenido: Vocabulario, relaciones verbales, lectura, cuantitativo, espacial
- Confiabilidad: puntuaciones totales, muy alta; subpuntuaciones, moderada (algunas, muy baja)
- Validez predictiva: por lo general, en el rango .30-.60
- Validez diferencial: en general, pobre
- Dos temas estadísticos especiales: restricción de rango y confiabilidad imperfecta
- Pruebas libres de cultura: hasta ahora, escurridizas

1. Contenido. A pesar de la diversidad de grupos meta y propósitos específicos de estas pruebas grupales, hay una notable semejanza en su contenido. Desde luego, difieren en el nivel de dificultad; sin embargo, teniendo en cuenta esto, por lo general encontramos reactivos de vocabulario, relaciones verbales, lectura, razonamiento cuantitativo y, en menor grado, razonamiento espacial.

2. Confiabilidad. Las puntuaciones totales en las pruebas grupales más usadas son, por lo común, más confiables. Los coeficientes de confiabilidad de consistencia interna suelen estar alrededor de .95, y los de test-retest, alrededor de .90. Las subpuntuaciones son menos confiables. Algunas pruebas proporcionan demasiadas subpuntuaciones basadas en muy pocos reactivos y sin niveles de confiabilidad suficientes para interpretarlas. Algunas personas piensan que las pruebas de aplicación grupal son, en general, menos confiables que las de aplicación individual, pero los datos no respaldan esta posición. En las pruebas elaboradas de manera adecuada, las puntuaciones totales de las pruebas grupales son al menos igual de confiables que las de las pruebas de aplicación individual. La ventaja principal de las pruebas individuales es la oportunidad de observar el desempeño del examinado de modo directo. Esto es especialmente importante cuando existe preocupación por una posible discapacidad, conducta desadaptada u otros padecimientos inusuales.

3. Validez predictiva. La validez predictiva de las pruebas grupales de capacidad mental es notablemente similar, aunque sin duda no idéntica, en una gran variedad de aplicaciones prácticas. Hemos examinado los criterios del desempeño en la escuela y el

trabajo. Es muy común que la validez predictiva de las puntuaciones totales se encuentre en un rango de .30 a .60.

Excepto al predecir el desempeño en otra prueba (donde las correlaciones son muy altas), la validez predictiva rara vez es mayor de .60. Por otro lado, rara vez es menor de .30.

Aquí hay mucho que pensar. La respuesta a la antigua pregunta de si la capacidad mental general hace alguna diferencia o es irrelevante en la vida es "sí". ¿La capacidad mental general hace alguna diferencia? Sí. ¿El esfuerzo, la determinación y el carácter hacen alguna diferencia? Sí. ¿Las circunstancias hacen alguna diferencia? Sí. ¿La suerte interviene en esto? Sí. ¿Hay error de medición? Sí, tanto de la prueba como del criterio. La vida es compleja, y también las relaciones entre estas variables lo es. Señalamos este punto en el capítulo sobre las pruebas grupales de capacidad mental, pero se aplica igual de bien a las pruebas individuales del capítulo anterior.

4. Validez diferencial. La esperanza nunca muere en relación con la validez diferencial de subpruebas en las pruebas grupales de capacidad mental. En general, varias combinaciones de subpruebas no producen coeficientes de validez superiores a los que se obtienen con las puntuaciones totales. Este punto tiene que ver explícitamente con el ASVAB. Una conclusión similar surge de un amplio rango de aplicaciones de pruebas en la selección laboral (Borman, Hanson, & Hedge, 1997; Kuncel & Hezlett, 2010; Schmidt, Ones, & Hunter, 1992). Lo mismo podría decirse de la admisión a universidades y programas de posgrado y de profesionalización. En general, una prueba relativamente breve pero confiable –digamos, de 50 reactivos con una confiabilidad de .90-.95– proporciona casi todo el poder predictivo de una batería de 3 hrs con múltiples subpruebas. Sin duda, es factible desarrollar una prueba de 50 reactivos para medir "*g*" con una confiabilidad de .90-.95; de hecho, no es tan difícil hacerla. Entonces, ¿por qué las personas siguen usando las baterías de 3 hrs y puntuaciones múltiples?

5. Restricción de rango y confiabilidad imperfecta. La investigación sobre las aplicaciones de las pruebas grupales de capacidad mental invariablemente implica algún tipo de situación predictiva.

En estas situaciones, necesitamos recordar el efecto de la falta de confiabilidad del criterio y, aún más, el de la restricción del rango. No es fácil recordar estos factores, pues no son evidentes para el sentido común. Los ajustes estadísticos no son sencillos, pero tampoco muy complicados y sí con consecuencias muy reales para nuestra comprensión de la manera en que funcionan las pruebas.

6. La búsqueda de una prueba libre de cultura no ha sido exitosa hasta ahora. Las pruebas consideradas como culturalmente neutrales tienden a ser pruebas no verbales de capacidad de razonamiento espacial y figurativo que no miden el mismo constructo o constructos que las típicas medidas de funcionamiento intelectual general.

Resumen

1. Las pruebas grupales de capacidad mental tienen sus usos más comunes en escuelas de todos los niveles, desde las escuelas primaras hasta las que ofrecen programas de posgrado y profesionalización, y en escenarios militares y de negocios. Su propósito primordial es, por lo general, hacer predicciones acerca del éxito en la escuela o el trabajo. Estas pruebas también se usan con fines de investigación.
2. Las pruebas grupales de capacidad mental tienen ocho características en común. La más obvia es que se aplican a grupos, aunque también se pueden aplicar de manera individual. Son de opción múltiple y se califican de manera automatizada. Su contenido, por lo general, es paralelo al de las pruebas individuales. Por lo común, tienen límites de tiempo y número de reactivos fijos, pero está aumentando la popularidad de las pruebas adaptada para computadora que no tienen estas características. Los tiempos de aplicación caen en dos bandos: los de 1 hr y los de 3 hrs. La mayoría de estas pruebas produce una puntuación total y varias subpuntuaciones. Tienden a tener fundamentos de investigación muy amplios. El principal propósito de casi todas las pruebas es la predicción.
3. A menudo se usa una prueba de capacidad mental de aplicación grupal junto con una prueba estandarizada de aprovechamiento en los programas de evaluación escolar. Describimos como ejemplo de estas pruebas la *Prueba de Capacidad Escolar Otis-Lennon*, Octava Edición, la más reciente de la larga línea de las pruebas Otis.
4. El SAT y el ACT se usan mucho como predictores del éxito en la universidad. Aunque tienen filosofías, historias y escalas de puntuación algo diferentes, estas pruebas son muy similares en su propósito, confiabilidad y validez.
5. Otro uso de las pruebas grupales de capacidad mental es seleccionar estudiantes en programas de posgrado y profesionalización. Describimos el *Graduate Record Examination: Prueba General* (GRE-G) como ejemplo de las pruebas de esta categoría.
6. Las pruebas grupales de capacidad mental también se usan para predecir el éxito en escenarios militares y de negocios. La *Batería de Aptitud Vocacional de las Fuerzas Armadas* (ASVAB) ilustra este tipo de uso. La estructura del ASVAB da un gran alivio al tema de usar varias subpruebas en este tipo de pruebas. En fuerte contraste, el Wonderlic es muy sencillo y corto.
7. Los psicólogos han buscado por mucho tiempo pruebas de capacidad mental culturalmente neutrales. Hasta la fecha, las pruebas que han resultado de estos esfuerzos son principalmente medidas de razonamiento figurativo/espacial. El muy citado *Matrices Progresivas de Raven* ofrece un buen ejemplo de estas pruebas.
8. Desarrollamos generalizaciones acerca de las pruebas grupales de capacidad mental en relación con su contenido, confiabilidad, validez predictiva y falta de validez diferencial. Pusimos especial atención en la necesidad de ser sensibles a dos cuestiones al usar pruebas para hacer predicciones: la restricción de rango y la confiabilidad

imperfecta del criterio.

Palabras clave

ACT
AFQT
ASVAB
College Board
ETS
GRE
OLSAT
prueba culturalmente neutral
prueba multinivel
Raven
reactivos de llenar óvalos
reactivo tipo matriz
restricción de rango
SAT
validez predictiva

Ejercicios

1. Para observar las diversas maneras en que las pruebas de inteligencia se pueden usar en la investigación, introduce "intelligence" como palabra clave en cualquier buscador de bases de datos electrónicas de artículos de ciencias sociales y de la conducta (p. ej., PsychINFO). Para evitar tener demasiadas referencias, limita la búsqueda a sólo uno o dos años. Trata de determinar con exactitud qué prueba se usó en algunos de los artículos que encontraste en la búsqueda.
2. Intenta con una prueba adaptadas para computadora; además de los temas cubiertos en este capítulo, entra a http://echo.edres.org:8080/scripts/cat/catdemo.htm. Realiza todo el ejemplo como si fueras "en verdad inteligente" y, después, "no tan inteligente".
3. Aprende a usar los sitios web para acceder a los datos técnicos de las pruebas. Como ejemplo, accede a la información del SAT en el sitio de College Board. Entra en www.collegeboard.com. Haz clic en Site Search y escribe en la casilla de Keyword "reliability". Examina algunos informes. Después, escribe "validity" como palabra clave y revisa algunos informes. También puedes acceder a los datos nacionales y estatales más recientes del SAT.
4. En la página 235a», presentamos algunos ejemplos de reactivos convertidos de respuesta libre en opción múltiple. Observa los reactivos muestra del cuadro 8-2. Trata de escribir tus propias versiones de opción múltiple de algunos de esos reactivos.
5. Observa el informe de las puntuaciones del OLSAT8 de la figura 9-3. ¿Cuál es el SAI del estudiante? ¿A qué conclusiones llegas acerca de este estudiante a partir del patrón de Comparaciones Anticipadas de Aprovechamiento?
6. En el ejercicio Inténtalo de la página 234a», hiciste una lista de las pruebas grupales de capacidad mental que has respondido; escoge una de ellas. Compárala con las ocho características comunes de las pruebas grupales de capacidad mental que se enumeran en la página 236a». ¿Qué tanto se ajusta la prueba que escogiste a estas ocho características?
7. Entra al sitio de *ACT Assessment*, www.act.org. ¿Cómo se describe el *propósito* del ACT en este sitio? ¿Puedes encontrar información sobre la *confiabilidad* de las pruebas ACT?
8. En la página 254a», señalamos las bajas confiabilidades de algunas pruebas del ASVAB. Por ejemplo, la prueba Comprensión de párrafos de 15 reactivos tiene una confiabilidad aproximada de .50. Empleando lo que has aprendido acerca de la relación entre confiabilidad y extensión de la prueba (véase p. 90c»), ¿cuántos reactivos debería tener esta prueba para alcanzar una confiabilidad aproximada de .85?
9. Usa los datos D1: GPA del apéndice D. Con el paquete estadístico de tu preferencia, prepara la distribución bivariada (dispersograma) del SAT Total frente al GPA. Además, obtén la correlación entre SAT y GPA. ¿A qué conclusiones llegas a partir de estos datos?

Notas

[1] Con fecha de agosto de 2011, una publicación importante del GRE (ETS, 2012) se refiere a la prueba como el "GRE revised General Test" y no sólo como "GRE General Test", distinción que con frecuencia se pasa por alto a lo largo del resto de la publicación y que nosotros también solemos pasar por alto.

[2] En 1977 se añadió la puntuación Analítica, que se convirtió en Escritura analítica en 2003. Antes de 1982, el GRE: Prueba General se conocía como GRE *Aptitud Test* [Prueba de aptitudes]. Un sorprendente número de fuentes aún usan este título y no distinguen la puntuación Analítica de la de Escritura analítica.

[3] Muy recientemente, la editorial cambió el nombre de la prueba de *Wonderlic Personnel Test* a *Wonderlic Cognitive Test* [Prueba Cognitiva Wonderlic], con ligeras variaciones en el nombre de las versiones en línea y de lápiz y papel. Aquí seguimos usando el nombre tradicional WPT, incluyendo las referencias al WPT-R.

CAPÍTULO 10
Evaluación neuropsicológica

Brooke J. Cannon
Matthew L. Eisenhard

Objetivos

1. Revisar la historia de la neuropsicología clínica.
2. Identificar las razones para realizar una evaluación neuropsicológica.
3. Describir los temas básicos de una evaluación neuropsicológica.
4. Distinguir entre un método de batería fija y uno de batería flexible.
5. Nombrar los dominios cognitivos que aborda el enfoque de batería flexible.
6. Enumerar las características de la seudodemencia.
7. Describir técnicas para evaluar esfuerzo/motivación.
8. Describir qué tipo de información complementaria debe reunirse cuando se hace una valoración neuropsicológica.
9. Nombrar dos tipos de dislexia y cómo difieren los errores en el deletreo y la lectura.

La evaluación neuropsicológica es uno de los sectores de más rápido desarrollo en el campo de las pruebas. En este capítulo, rastreamos los orígenes de este tipo de evaluación y describimos los tipos de problemas que trata. Encontraremos que se sirve de muchos conceptos y pruebas específicas que hemos visto en capítulos anteriores. De hecho, veremos qué tan importante es conocer el material anterior para poder llevar a cabo una evaluación neuropsicológica competente. Empecemos con tres casos que demandan evaluación neuropsicológica. Concluiremos el capítulo viendo qué

evaluaciones tuvieron lugar efectivamente en cada caso.

Casos

- Caso #1. Nancy O'Roarke, mujer de 74 años de edad que recientemente enviudó, se queja de pérdida de memoria, problemas para concentrarse e incapacidad para realizar sus actividades cotidianas. ¿Tiene la enfermedad de Alzheimer?
- Caso #2. Jack Davis, trabajador de la construcción de 42 años de edad, sufrió una lesión en su trabajo cuando un bloque de cemento lo golpeó en la cabeza. Desde entonces, tiene dolores de cabeza, pérdida de memoria, períodos breves de atención y un estado de ánimo deprimido. Se encuentra en un pleito legal contra su patrón. ¿Tiene daño cerebral?
- Caso #3. Billy Zollinger, niño de 10 años de edad de cuarto grado, ha bajado sus notas en la escuela. Su madre informa que a menudo pierde cosas, olvida hacer lo que ella le pide o no logra terminar sus trabajos. Además, le toma un tiempo excesivo hacer sus tareas escolares, y a menudo se distrae con ruidos exteriores y otros estímulos. ¿Tiene trastornos por déficit de atención?

El cerebro: camino a la neuropsicología clínica

Hoy es difícil imaginar que el cerebro no siempre se consideró como el sitio donde se encuentra la mente. Los antiguos egipcios pensaban que los intestinos, o vísceras, eran los responsables del pensamiento, por lo que les daban un tratamiento cuidadoso en las momias para que pudieran ser usados en la vida del más allá, mientras que ¡desechaban el cerebro! El primer registro que se tiene de la "hipótesis del cerebro" se atribuye a Alcmeón de Crotona (ca. 500 a. de C.), quien propuso que el cerebro controlaba las capacidades mentales. Una conciencia más específica de la relación entre el cerebro humano y la capacidad para razonar apareció hace cerca de 2000 años, cuando el médico romano **Galeno** (129-ca. 210 d. de C.) atendía a los gladiadores heridos. Por medio de su trabajo y su exposición a distintos tipos de heridas, determinó que el cerebro era un órgano crítico, responsable de los sentidos, el lenguaje y el pensamiento.

Una vez que se conoció la relación entre cerebro y conducta, se hicieron intentos por comprender en qué partes del cerebro residían las distintas capacidades. **Franz Josef Gall** (1758-1828) desarrolló el concepto de **frenología**, el estudio de las relaciones entre las conductas morales, emocionales e intelectuales y las variaciones de la superficie del cráneo (figura 10-1). Aunque la frenología no resistió un examen minucioso, Gall inició la búsqueda de la localización de las funciones en el cerebro.

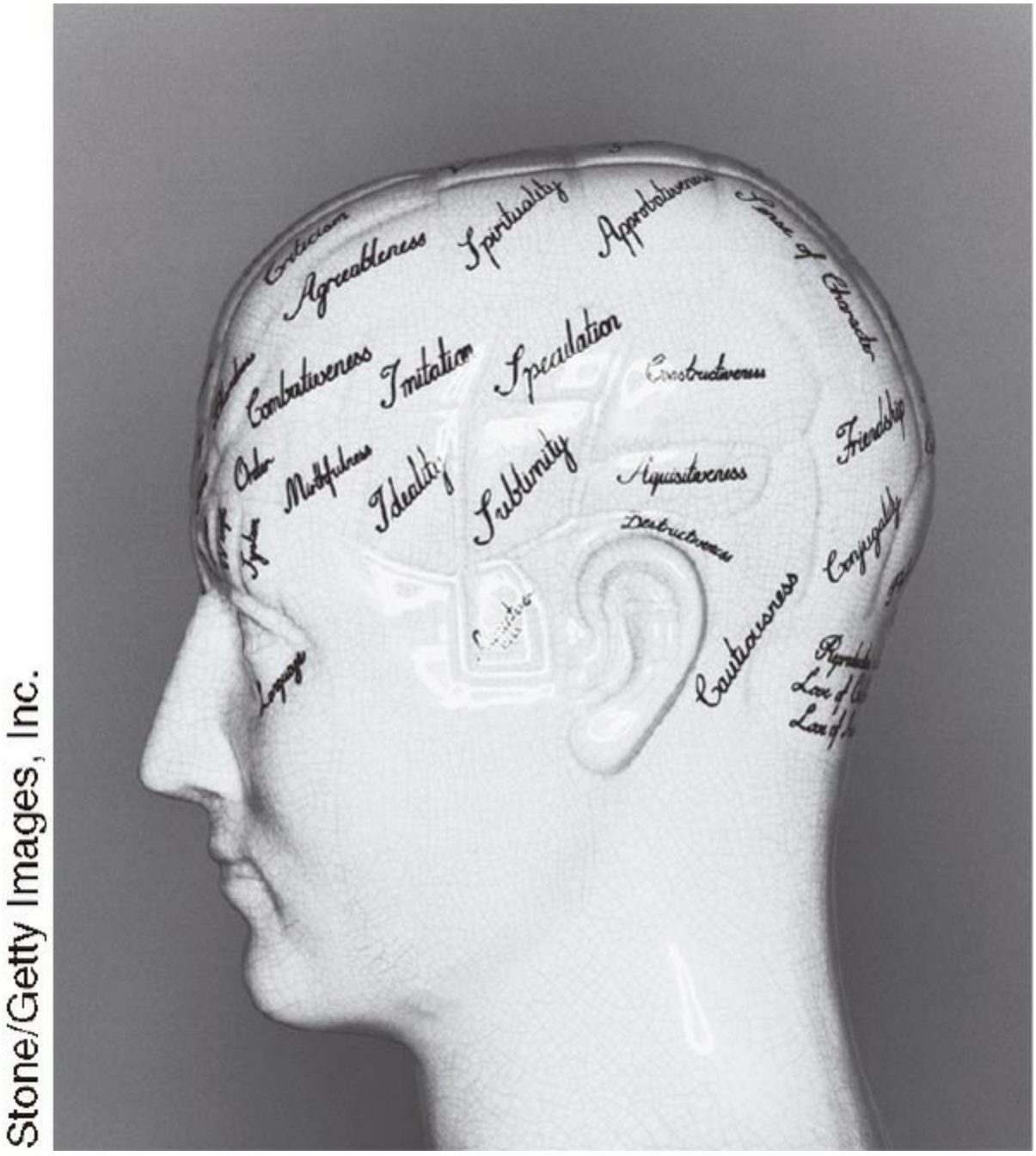


Figura 10-1. Áreas del cráneo de acuerdo con la teoría frenológica.

La segunda mitad del siglo XIX fue testigo de avances importantes en la localización de las capacidades de lenguaje. Mediante un profundo estudio de caso, **Paul Broca** (1824-1880), cirujano francés, fue el primero en documentar el sitio del daño cerebral asociado con la incapacidad para hablar, pero con la comprensión del lenguaje intacta. **Carl Wernicke** (1848-1904), neuroanatomista alemán, describió una perturbación del lenguaje que implicaba comprensión afectada, pero con el habla intacta, aunque no era muy importante la afectación. Se encontró que este segundo trastorno de lenguaje estaba

asociado con un área diferente del cerebro. Los problemas de comunicación, como los descritos por Broca y Wernicke, se conocen como **afasias**, déficit en la capacidad para expresar o comprender el lenguaje escrito o hablado como resultado de una lesión cerebral.

Uno de los estudios de caso más famosos de cambios de personalidad después de una lesión cerebral es el de Phineas Gage, un trabajador del ferrocarril que, en 1848, sufrió daño cerebral cuando una explosión disparó una barra de hierro que atravesó su cabeza (figura 10-2).. La barra, de un metro de largo y tres centímetros de diámetro entró por su mejilla izquierda y salió por la parte superior de su cabeza, de modo que dañó la porción frontal de su cerebro. Antes del accidente, Gage era responsable, eficiente y hábil para los negocios, pero después de la lesión, era odioso e impaciente (Macmillan, 2000). Empezó a usar un lenguaje profano y no podía seguir ningún plan de acción que él mismo ideaba.

Figura 10-2. Phineas Gage, cuya lesión cerebral condujo a cambios de personalidad. *Fuente:* M. B. MacMillan. *Brain and Cognition*, 1986, p. 85. Reproducido con autorización de Academic Press.

Al principio, el campo de la **neuropsicología**, definido como el estudio de las relaciones entre cerebro y conducta, correspondía a los psicólogos fisiológicos, quienes trabajaban primordialmente con modelos animales del funcionamiento cerebral. A Arthur Benton (1909-2006) se puede atribuir el nacimiento de la **neuropsicología clínica**, una especialidad profesional distinta que combina la neuropsicología humana con la psicología clínica. El neuropsicólogo clínico aplica los principios cerebro-conducta en la evaluación y tratamiento de sus pacientes. Benton empezó a formar estudiantes de doctorado en neuropsicología clínica en la Universidad de Iowa, y las primeras dos tesis de esta disciplina se terminaron en 1954 (Benton, 1997). El cuadro 10-1 presenta una línea del tiempo con los eventos importantes de la historia de la neuropsicología clínica.

Cuadro 10-1. Eventos clave en el desarrollo de la neuropsicología

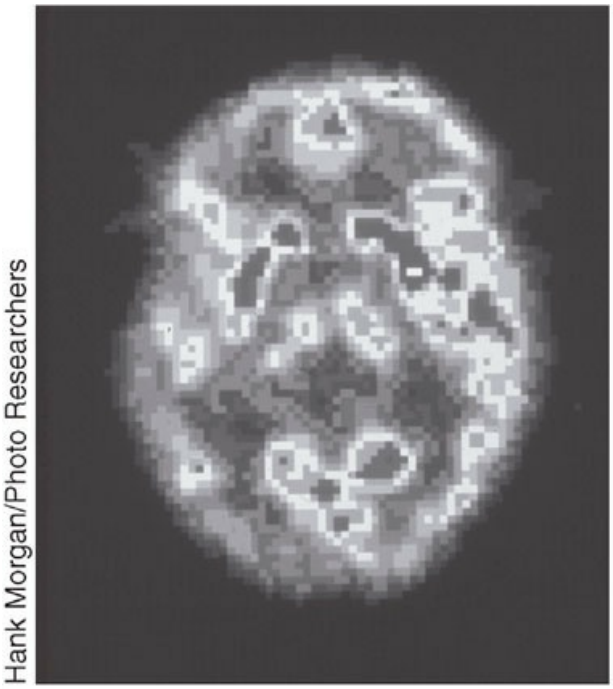
500 a. de C.	Alcmeón de Crotona propone que las capacidades mentales son controladas por el cerebro.
~180 d. de C.	Galeno determina que el cerebro es crítico para la sensibilidad, el lenguaje y el pensamiento.
1798	Gall desarrolla la frenología.
1848	Phineas Gage: una barra de metal atraviesa su cabeza.
1861	Broca informa el sitio de un daño cerebral asociado con déficit en el lenguaje expresivo.
1874	Wernicke descubre el sitio de un daño cerebral asociado con la comprensión del lenguaje.
1954	Benton supervisa las primeras tesis doctorales en neuropsicología clínica.
1967	Se crea la Sociedad Internacional de Neuropsicología.
1979	Se establece APA División 40, Neuropsicología Clínica.
1996	La APA reconoce oficialmente a la neuropsicología clínica como una especialidad.

Diagnósticos más frecuentes en la evaluación neuropsicológica	
Adultos	*Niños*
Demencia senil	TDAH
Lesión cerebral por traumatismo craneal cerrado	Problemas de aprendizaje
Derrame cerebral o accidente cardiovascular	Trastornos convulsivos
Otros padecimientos médicos/neurológicos	Lesión cerebral por traumatismo craneal cerrado
TDAH	Trastornos generalizados del desarrollo

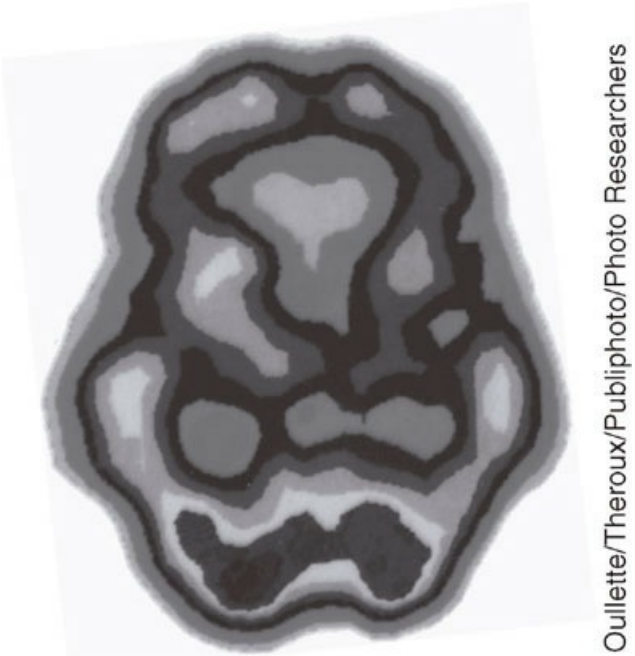
Fuente: Sweet, Meyer, Nelson y Moberg, 2011.

Se acude al neuropsicólogo clínico para responder a muchas preguntas acerca de las personas que se envían con dicho especialista. Podemos identificar seis razones importantes de la evaluación neuropsicológica. Primero, la evaluación se puede solicitar para establecer un *diagnóstico*. Una encuesta reciente (Sweet, Meyer, Nelson, & Moberg, 2011) señaló los siguientes diagnósticos de padecimientos encontrados en las evaluaciones neuropsicológicas pediátricas y de adultos.

Las pruebas neuropsicológicas pueden demostrar una discapacidad cognitiva cuando otros exámenes médicos (p. ej., rayos X de la cabeza) arrojan resultados normales. Con la evolución de las técnicas de escaneado del cerebro, la pregunta general "¿hay daño cerebral?" ya no es para el neuropsicólogo clínico. La introducción de la tomografía axial computarizada (TAC), o escáner, permite hacer un examen rápido y detallado del cerebro. La imagen por resonancia magnética (IRM) brinda una claridad aún mayor de las estructuras cerebrales, en especial de los pequeños cambios causados por mala circulación sanguínea en el cerebro. En años más recientes, la tomografía por emisión de positrones (TEP) y la menos cara tomografía computarizada de emisión monofotónica (SPECT, siglas en inglés) abren una ventana a la actividad del cerebro en vez de las simples fotos de sus estructuras (véase figura 10-3). Con estos avances tecnológicos es posible ver qué área del cerebro se usa durante distintas actividades. Por medio del estudio del funcionamiento cerebral normal, es posible determinar qué áreas del cerebro están menos activas en diferentes enfermedades o lesiones. Aunque estas tecnologías pueden demostrar que hay un daño en cierta estructura del cerebro o que hay una actividad metabólica por debajo de lo normal, pueden no ser capaces de decirnos exactamente cómo es la conducta del paciente. Con sólo mirar una TAC o una IRM del cerebro, no se puede decir si el paciente tiene algún tipo de conducta, pues ¡podría estar muerto!



Hank Morgan/Photo Researchers



Oullette/Theroux/Publiphoto/Photo Researchers

Figura 10-3. Imágenes de un TEP y SPECT.

La evaluación neuropsicológica ya no responde a la pregunta de "si hay daño cerebral o no"; sin embargo, aún debe ser ella la que responda a las preguntas acerca del diagnóstico y no las imágenes del cerebro. Por ejemplo, en el caso de la pregunta "¿este paciente tiene demencia?", la TAC o la IRM son normales en los inicios de la demencia, pero los resultados de la evaluación neuropsicológica indican la presencia de anormalidades cognitivas. Una persona puede presentarse sin déficit en un examen médico superficial, pero una evaluación cognitiva detallada podría descubrir déficit. La evaluación neuropsicológica es valiosa para documentar este déficit. Recomendamos consultar *Guidelines for the Evaluation of Dementia and Cognitive Chang*e [Directrices para la evaluación de la demencia y el cambio cognitivo] de la American Psychological Association (2012).

En muchas ocasiones, el diagnóstico ya se sabe, como un derrame cerebral confirmado por el examen neurológico y el escaneado del cerebro. Ahora la pregunta que se plantea al neuropsicólogo clínico es: "¿cuáles son las *f*ortalezas y debilidades del paciente después de sufrir una lesión cerebral?" Aquí, la evaluación neuropsicológica puede documentar áreas donde el funcionamiento se preservó y áreas de deterioro. Este conocimiento es útil para planear la formación profesional cuando el paciente intenta regresar al trabajo o decide continuar con su educación. Alguien que ha sufrido un derrame cerebral que le produjo una pobre atención visual ¡no debería tener un cargo de control del tráfico aéreo!

Mientras el paciente está en el hospital, se puede pedir al neuropsicólogo clínico ayudar en la planeación del tratamiento. Por ejemplo, la evaluación puede demostrar que el paciente recuerda información cuando es de naturaleza verbal; un paciente puede ser impulsivo o tener déficit de memoria que le impediría vivir de manera independiente. El neuropsicólogo clínico puede usar los resultados de la evaluación para ayudar al personal del hospital en el manejo del paciente, pues es posible que no todos los encargados del tratamiento estén conscientes de los posibles cambios conductuales asociados con las lesiones cerebrales. Por ejemplo, una enfermera estaba muy enojada porque un paciente la maldecía constantemente y nunca dijo algo agradable. Resultó que el paciente tenía una afasia debida al derrame cerebral y las únicas palabras que aún podía articular eran las de maldecir. Quizá trataba de decir "buenos días", pero en lugar de eso salía una maldición. Cuando la enfermera se dio cuenta de la asociación entre la conducta del paciente y su lesión cerebral, aprendió a observar su comunicación no verbal e ignorar su vocabulario.

Otra área donde se ha utilizado la evaluación neuropsicológica es la forense. Los neuropsicólogos clínicos pueden ser contratados por los abogados defensores o demandantes o por la corte. A menudo se les pide que determinen si existe algún déficit cognitivo y si concuerda con la lesión cerebral que sufrió el cliente. Por último, la evaluación neuropsicológica se usa en la investigación para estudiar el funcionamiento cognitivo normal, así como la conducta de alguien de quien se sabe o se supone tiene

deterioro cerebral. La evaluación también puede ayudar a determinar los efectos cognitivos del tratamiento médico o de un nuevo medicamento.

Resumen de puntos clave 10-1

Razones para la evaluación neuropsicológica
Diagnóstico
Identificar fortalezas y debilidades
Planeación de la formación profesional
Planeación del tratamiento
Área forense
Investigación

Dos métodos de evaluación neuropsicológica

Existen dos métodos importantes de evaluación neuropsicológica. El primero es la **batería fija**, en la que se usan las mismas pruebas (una batería) con todos los examinados; la batería consta de muchas subpruebas. El segundo método se denomina **batería flexible**, la cual permite al clínico elegir las subpruebas que cree más adecuadas para un examinado. Es importante recordar que las subpruebas en ambos métodos son autocontenidas. Las pruebas individuales de una batería fija (p. ej., *Trail Making Test*) se pueden usar como parte de una batería flexible.

Método de batería fija

Cerca de 5% de neuropsicólogos clínicos usa una batería fija estandarizada en la evaluación neuropsicológica y 18% usa una batería flexible, en la cual puede incorporarse a veces la evaluación con baterías fijas (Sweet *et al*., 2011). Las dos baterías que más se usan son el **Luria-Nebraska Neuropsychological Battery** [Batería Neuropsicológica Luria-Nebraska] y el **Halstead-Reitan Neuropsychological Battery** [Batería Neuropsicológica Halstead-Reitan].

Batería Neuropsicológica Luria-Nebraska

La Batería Neuropsicológica Luria-Nebraska (BNLN; Golden, Purisch, & Hammeke, 1985) surgió de la "Investigación Neuropsicológica de Luria" (Christensen, 1984), un conjunto de análisis cualitativos de la conducta de pacientes basado en la obra del neuropsicólogo ruso Aleksandr R. Luria (1902-1977). La BNLN se modificó para producir puntuaciones de 11 escalas clínicas, dos sensoriomotrices, seis de localización adicional y cinco de resumen (cuadro 10-2). Veintiocho escalas de factores adicionales permiten determinar el funcionamiento sensorial y cognitivo más específico. Existen básicamente dos formas equivalentes de la prueba, pero la Forma II incluye una escala clínica adicional. Esta prueba está diseñada para individuos de 15 años de edad en adelante. También hay una forma para niños de 8 a 12 años que incluye las mismas escalas clínicas y opcionales de la forma para adultos, pero no cuenta con escalas de localización y sólo tiene tres de resumen. Hay 11 escalas de factores adicionales en la versión para niños.

Cuadro 10-2. Escalas de la forma para adultos de la Batería Neuropsicológica Luria-Nebraska

Escalas clínicas	Escalas de localización	Escalas de resumen	Escalas opcionales
Aritmética	Frontal izquierda	Hemisferio izquierdo	Motriz
Lenguaje expresivo	Sensoriomotriz izquierda	Hemisferio derecho	Escritura
Procesos intelectuales	Parieto-occipital izquierda	Patognomónico	Deletreo
Memoria intermedia	Temporal izquierda	Elevación del perfil	
(sólo Forma II)	Frontal derecha	Deterioro	
Memoria	Sensoriomotriz derecha		
Funciones motrices	Parieto-occipital derecha		
Lectura	Temporal derecha		
Lenguaje receptivo			
Ritmo			
Funciones táctiles			
Funciones visuales			
Escritura			

La aplicación de la BNLN requiere de 1.5 a 2.5 hrs, y puede calificarse a mano o por computadora. La batería de pruebas es portátil, por lo que se puede aplicar en la cama si es necesario.

La validación de la BNLN se concentra en diferenciar entre los pacientes con daño cerebral y los de otros grupos. Parece distinguir con bastante exactitud los pacientes con daño cerebral de los individuos normales. Golden y otros autores de pruebas han presentado varios informes de la exactitud diagnóstica de la BNLN (p. ej., Golden, Hammeke, & Purisch, 1979; Golden, 1981, 1984). Sin embargo, diversos estudios contradicen estos resultados (p. ej., Adams, 1980, 1984; Crosson & Warren, 1982). Las críticas a la BNLN incluyen su incapacidad para detectar las discapacidades moderadas y la inexactitud de las escalas de localización; además, se ha encontrado que las escalas de memoria son afectadas por déficit de atención, de modo que no se puede concluir claramente que existe un deterioro de la memoria. Asimismo, debido a que muchos reactivos requieren de procesamiento verbal, la presencia de un déficit de lenguaje puede crear un sesgo en la prueba (Franzen, 2000).

Las ventajas de la BNLN incluyen su facilidad de aplicación, su naturaleza portátil y su brevedad; sin embargo, más allá de confirmar la presencia de daño cerebral, su utilidad es cuestionable. Aunque la BNLN tiene partidarios fervientes, no es tan popular como otras evaluaciones neuropsicológicas. Una revisión de 100 evaluaciones neuropsicológicas realizadas con propósitos forenses encontró que en sólo 10% se usó el BNLN (Lees-Haley, Smith, Williams, & Dunn, 1996).

Batería de Pruebas Neuropsicológicas Halstead-Reitan

La **Batería Neuropsicológica Halstead-Reitan** (BNHR; Reitan & Wolfson, 1993) se usa con adultos, pero también hay una versión para niños de 9 a 14 años de edad. Esta batería consta de 10 pruebas (cuadro 10-3); el desempeño en cinco de ellas determina el Índice de deterioro (Prueba de categorías, Prueba de desempeño táctil, Prueba de ritmos de la playa, Prueba de percepción de sonidos del habla, Prueba de golpeteo dactilar). Este índice proporciona un punto de corte para representar la presencia o ausencia de déficits neurológicos. Por lo general, también se aplican la Escala Wechsler de Inteligencia para Adultos y el Inventario Multifásico de Personalidad de Minnesota, pero no contribuyen al Índice de deterioro. También es necesaria la evaluación de la memoria, ya que no hay una prueba de memoria en la BNHR. Se emplea un número mayor de variables (42) para obtener la **Puntuación de Déficit Neuropsicológico General**, que refleja la gravedad del déficit neuropsicológico. Es decir, el Índice de deterioro se usa para determinar la presencia de un déficit y la Puntuación de déficit neuropsicológico refleja el grado de deterioro.

Varios estudios sobre la BNHR y sus índices han demostrado su capacidad para discriminar sujetos con daño cerebral de sujetos normales (Russell, 1995), con un índice general de exactitud de 80%. También se ha encontrado que la BNHR tiene una buena confiabilidad de test-retest (Russell, 1992). Algunos inconvenientes de esta batería son la

falta de una prueba de memoria y su tiempo de aplicación (de 4 a 5 hrs para un examinador experimentado). Originalmente, la batería estaba diseñada para determinar la presencia o ausencia de deterioro, como se refleja en el Índice de deterioro. De hecho, la investigación empírica de los autores de una de las pruebas refleja su utilidad como "indicador general de las funciones cerebrales" (Reitan & Wolfson, 1989) y no para discriminaciones más finas. Otros han encontrado una falta de contribución de un diagnóstico único de algunas pruebas de la batería (Sherer, Parsons, Nixon, & Adams, 1991).

Método de batería flexible

En contraste con las baterías fijas descritas antes, las flexibles permiten al neuropsicólogo clínico elegir las pruebas de acuerdo con el paciente y el motivo de consulta. Adoptar este método excluye el uso del índice de deterioro que se obtiene de las baterías fijas. Conforme la neuropsicología se ha alejado de la cuestión "daño cerebral o no", la batería flexible se ha adoptado para ajustar la evaluación a las razones para llevarla a cabo y dar una descripción más detallada de los déficits que están presentes. En una encuesta reciente (Sweet *et al*., 2011) se encontró que 78% de los neuropsicólogos clínicos usan el método de batería flexible, mientras que sólo 5% se basan en el de batería fija.

La selección de pruebas, por lo común, sigue una evaluación planeada de varios dominios cognitivos. El cuadro 10-4 contiene ejemplos de pruebas usadas en cada uno de estos dominios. Discutiremos algunas de éstas más tarde. El cuadro 10-5 enumera las pruebas que se usan con mayor frecuencia en la evaluación neuropsicológica. Podemos notar que algunas de éstas, por ejemplo, las escalas Wechsler y el MMPI, se abordan en otros capítulos y se usan en contextos como complemento de la evaluación neuropsicológica. En Strauss, Sherman y Spreen (2006) se pueden encontrar descripciones detalladas de los procedimientos de aplicación, así como datos normativos de varias fuentes de muchas baterías flexibles que se mencionan más adelante.

Cuadro 10-4. Ejemplos de pruebas empleadas para evaluar diversos dominios cognitivos

Atención	**Memoria**
Continuous Performance Test	*Auditory Consonant Trigrams*
Paced Auditory Serial Addition Test	*Autobiographical Memory Interview*
Symbol Digit Modality Test	*Benton Visual Retention Test*
Trail-Making Test	*Buschke Selective Reminding Test*
Aprovechamiento	*California Verbal Learning Test*
Peabody Individual Achievement Test	*Rey Auditory Verbal Learning Test*
Wide Range Achievement Test	*Wechsler Memory Scale*
Wechsler Individual Achievement Test	**Estado mental**
Esfuerzo/Motivación	*Mini-Mental State Exam*
Rey Fifteen-Item Memory Test	**Motricidad**
Test of Memory Malingering	*Finger-Tapping Test*
21-Item Test	*Grooved Pegboard Test*
Validity Indicator Profile	*Hand Dynamometer*
Victoria Symptom Validity Test	*Purdue Pegboard Test*
Funciones ejecutivas	**Personalidad/Estado psicológico**
Behavioral Assessment of the Dysexecutive System	*Beck Depression Inventory*
California Sorting Test	*Child Behavior Checklist*
Category Test	*Geriatric Depression Scale*
Delis Kaplan Executive Function System	*Minnesota Multiphasic Personality Inventory*
Design Fluency Test	*Neurobehavioral Rating Scale*
Stroop Test	*Neuropsychology Behavior and Affect Profile*

Wisconsin Card Sorting Test	*Profile of Mood States*
Inteligencia	*Vineland Adaptive Behavior Scales*
Kaufman Brief Intelligence Test	**Capacidad visoespacial/perceptual**
Microcog: Assessment of Cognitive Functioning	*Clock Drawing*
Raven's Progressive Matrices	*Embedded Figures Test*
Wechsler Intelligence Scales	*Facial Recognition Test*
Lenguaje	*Hooper Visual Organization Test*
Boston Naming Test	*Rey/Osterrieth Complex Figure Test*
Controlled Word Association	*Right-Left Orientation*
Peabody Picture Vocabulary Test	
Token Test	

Cuadro 10-5. Las 10 pruebas más usadas por los neuropsicólogos

Lugar	Prueba
1.	Inventario Multifásico de Personalidad de Minnesota
2.	Escala Wechsler de Inteligencia para Adultos
3.	Escala de Memoria de Wechsler – Revisada
4.	Prueba de trazo, A y B
5.	Prueba de asociación oral controlada de palabras
6.	Prueba de golpeteo dactilar
7.	Batería de Pruebas Neuropsicológicas Halstead-Reitan
8.	Prueba Boston de nombres
9.	Prueba de categorías
10.	Prueba de aprovechamiento de rango amplio

Fuente: Camara, Nathan y Puente (2000).

Estado mental

Cuando se considera qué pruebas aplicar a un paciente, primero es necesario obtener una imagen general de su **funcionamiento cognitivo grueso.** Si el paciente tiene demencia, por ejemplo, no sería deseable aplicar pruebas demasiado difíciles, pues sería una pérdida de tiempo para el paciente y el examinador. Es mejor empezar explorando en busca de deterioros cognitivos importantes. La medida que se usa con mayor frecuencia para este propósito es el *Mini Examen del Estado Mental* (MEEM), que consta de 30 puntos. Los reactivos miden orientación general (p. ej., "¿Qué día es hoy?"), memoria (p. ej., "Recuerde estas tres palabras: MESA, PEZ, CAJA"), lenguaje (p. ej., tomar un lápiz y preguntar al paciente cómo se llama el objeto), control mental (p. ej., "deletrea 'mundo' en orden inverso"), habilidades visoconstructivas (copiar un diseño geométrico) y la capacidad para seguir una orden de pasos múltiples. Los reactivos son intencionalmente fáciles; de los 30 puntos, un adulto normal tiene 29 o 30, por lo que las puntuaciones menores de 24 se consideran un indicio de deterioro significativo en el funcionamiento

cognitivo grueso. Sin embargo, una calificación de 29 o 30 no significa que el paciente está por completo libre de deterioro.

Resumen de puntos clave 10-2

Dos métodos importantes de evaluación neuropsicológica

1. Método de batería fija
2. Método de batería flexible

La edición original de esta venerable medida (Folstein, Folstein, & McHugh, 1975; Folstein, Folstein, McHugh, & Fanjiang, 2000) ahora está disponible en tres versiones, como se describe más adelante (MEEM-2; Folstein, Folstein, White, & Messer, 2010). La versión estándar se considera comparable con la prueba original con algunos ajustes menores. La versión breve (mini-mini) contiene cerca de la mitad de los puntos de la versión estándar y su aplicación requiere apenas 5 min. La versión *extensa* (maxi-mini) incluye los 30 puntos de la versión estándar más otros 60 puntos con material de mayor dificultad que busca identificar deterioro cognitivo menos grave. ¿Cuál versión resultará más popular? Sólo el tiempo lo dirá; mientras tanto, si alguien dice que usó el MEEM, habrá que preguntar qué versión usó.

Inteligencia

Las escalas de inteligencia Wechsler casi siempre se usan como medidas de inteligencia. Como ya las vimos en el capítulo 8, no las describiremos aquí. En la versión más reciente, el WAIS-IV, el neuropsicólogo clínico considera las puntuaciones de los cuatro índices del CI, así como el desempeño en las subpruebas individuales. Una diferencia grande entre los índices de Comprensión verbal y Razonamiento perceptual, por ejemplo, puede reflejar un deterioro relativo en el lenguaje o en el funcionamiento no verbal. Los índices de Memoria de trabajo y Velocidad de procesamiento añaden información respecto de los deterioros en el procesamiento cognitivo. También se determinan las fortalezas y debilidades en las 10 subpruebas principales y en las cinco suplementarias. Las puntuaciones de las subpruebas corregidas por edad se promedian y el neuropsicólogo clínico busca desviaciones de tres puntos o más respecto del promedio en cada subprueba. Las estimaciones de CI **premórbidos** (antes del inicio de un deterioro) se pueden determinar por medio de fórmulas basadas en factores como el sexo, ocupación y aprovechamiento educativo del paciente. Los registros educativos contienen a menudo estimaciones de CI premórbidos, ya que muchos niños realizan una prueba de inteligencia en la escuela.

Aprovechamiento

Otra indicación de un nivel cognitivo premórbido puede ser la capacidad para leer, deletrear y resolver problemas aritméticos. A menudo, la Prueba de Aprovechamiento de Rango Amplio (WRAT, siglas en inglés) se usa para estimar el aprovechamiento académico con subpruebas que evalúan estas tres áreas (con subpruebas separadas para lectura de palabras y comprensión de lectura; el WRAT también proporciona una puntuación compuesta de lectura que toma en cuenta el desempeño en ambas subpruebas). El desempeño del paciente también proporciona información acerca de posibles problemas de aprendizaje; por ejemplo, errores en la lectura o el deletreo pueden reflejar dislexia. Errores en los problemas matemáticos, como sumar en vez de restar, pueden estar relacionados con déficit de atención. El desempeño del paciente en el WRAT se informa en percentiles y equivalentes de grado.

Atención/concentración

El método de batería flexible incluye varias pruebas con componentes de atención/concentración. Se considera el desempeño en las subpruebas del WAIS; por ejemplo, Retención de dígitos es una medida de atención auditiva. A menudo, los pacientes son capaces de repetir dígitos en el orden en que los escucharon sin dificultades, pero cuando se les pide que los repitan en orden inverso, se hacen evidentes los problemas con la concentración y el control mental. Pueden presentarse problemas similares en la subprueba Sucesión de números y letras. Los déficits de atención visual pueden reflejarse en puntuaciones bajas en Búsqueda de símbolos, Registros o Figuras incompletas.

La Prueba de trazo es un componente de la Batería Neuropsicológica Halstead-Reitan que se aplica a menudo como parte de una batería flexible. Esta prueba es una medida de exploración visual, velocidad de escritura y atención/concentración. En el cuadro 10-3 se encuentra una descripción de la prueba. La parte más difícil, B, requiere unir círculos en orden mientras se alternan número-letra-número-letra. De esta manera, las respuestas correctas serían 1-A-2-B-3-C y así sucesivamente. Los pacientes con lesión cerebral pueden pasar un momento muy difícil al realizar esta tarea. A menudo **perseveran** o son incapaces de cambiar sus patrones de pensamiento.

Cuadro 10-3. Componentes de la Batería Neuropsicológica Halstead-Reitan

Prueba de exploración de afasias	El paciente debe realizar diferentes tareas sencillas: nombrar, deletrear, escribir y leer, identificar partes del cuerpo; aritmética simple, diferenciación izquierda-derecha y copia de formas geométricas sencillas. La hipótesis es que una persona normal podría realizar estos reactivos con facilidad; los errores indican una disfunción cerebral.
Prueba de categorías	El paciente responde a una serie de figuras que aparecen muy brevemente en una pantalla (o en un cuadernillo de estímulos) presionando una de cuatro palancas. Una campana (correcto) o un timbre (incorrecto) suenan después de cada respuesta. Cada conjunto de estímulos tiene un principio unificador, el cual el paciente tiene que determinar por medio de la retroalimentación que recibe antes de sus respuestas.
Prueba de golpeteo	Ésta es una medida de velocidad motriz. El paciente tiene que golpetear el aparato con su dedo índice tan rápido como pueda durante ensayos de 10 segundos. Esto se hace

dactilar	tanto con la mano dominante como con la mano no dominante.
Fuerza de prensión	El aparato que se usa para esta prueba es el "dinamómetro manual", el cual el paciente sujeta con el brazo extendido lateralmente y aprieta el puño. De nuevo, esto se hace con las dos manos.
Prueba de ritmo	Usando una cinta de audio, se presenta al paciente varios pares de redobles rítmicos que tiene que identificar como iguales o diferentes.
Examen perceptual-sensorial	Esta evaluación incluye una sub-batería de medidas que evalúan la percepción del sonido, tacto y visión. En particular, hay estímulos unilaterales que ocurren en un lado del cuerpo y otros que son bilaterales. Los pacientes con ciertos tipos de daño cerebral pueden desempeñarse bien con estímulos unilaterales, pero lo hacen mal con los bilaterales.
Prueba de percepción de sonidos del habla	Usando de nuevo una cinta de audio, se presentan 60 palabras habladas sin sentido, las cuales el paciente tiene que identificar entre cuatro opciones en la hoja de respuestas.
Prueba de reconocimiento de formas táctiles	Aquí, el paciente tiene que identificar por medio del tacto varias piezas de plástico de diferentes formas, como un cuadrado o un triángulo.
Prueba de desempeño táctil	Para esta prueba, se vendan los ojos al paciente antes de presentarle un tablero vertical con varias formas recortadas a las que les corresponde un bloque. Primero con una mano y luego con ambas, el paciente tiene que colocar el bloque correcto en el espacio correspondiente. Después se esconden los materiales y se le retira la venda de los ojos. Entonces se le pide que dibuje el tablero con las formas en los lugares correctos.
Prueba de trazo	Hay dos componentes en esta prueba, Parte A y Parte B. Cada una se presenta en una hoja de papel de tamaño estándar. La Parte A incluye círculos que contienen números hasta el 25. El paciente tiene que unir los círculos en orden con una línea continua tan rápido como sea posible. La Parte B es similar, pero contiene números del 1 al 13 y letras de la A a la L. Tiene que unir de nuevo estos círculos en orden, pero esta vez debe alternar entre números y letras (1-A-2-B).

¡Inténtalo!

Pide a un amigo que te lea las siguientes secuencias de números a una velocidad de un número por segundo, empezando por las más cortas hasta las más largas. Después de que te lea cada serie, repite los dígitos en el mismo orden. ¿Cuál es la serie de dígitos más larga que pudiste recordar sin errores? Ahora pide a tu amigo que te lea cada serie otra vez, pero ahora repite los dígitos en orden inverso.

4-2-7
6-9-8-3
2-1-7-4-8
8-6-2-5-9-7
9-3-7-5-6-2-4
7-1-4-2-8-3-6-9-5

¿Qué fue más difícil? La mayoría de las personas puede recordar 7 +/−2 dígitos en el mismo orden en que los escucha y cerca de dos dígitos menos cuando los debe recordar en orden inverso.

Lenguaje

El deterioro en la comunicación puede ser obvio, como cuando un paciente tiene afasia o un déficit más sutil. El lenguaje implica capacidades tanto expresivas como receptivas; hablar y escribir son habilidades expresivas, mientras que comprender y leer son habilidades receptivas. Un ejemplo de una medida de lenguaje expresivo es la Prueba Boston de nombres, la cual comprende 60 dibujos de objetos ordenados del más fácil al más difícil, es decir, el primer reactivo podría ser tan fácil como "silla" y el último podría ser tan difícil como "remache" (tú sabes, esas pequeñas cosas de metal que hay en los bolsillos de tus *jeans*). Una persona con un deterioro grave en la capacidad para nombrar puede ser incapaz de decir el nombre de un dibujo de manera espontánea. Sin embargo, si se le dice una pista fonémica (p. ej., "empieza con 'ch'"), la persona puede responder correctamente.

Otra prueba de lenguaje expresivo puede realizarse de manera oral o escrita; se trata de la Prueba de asociación controlada de palabras, también conocida como fluidez de palabras o fluidez verbal. Aquí, la tarea es enumerar tantas palabras como sea posible que empiecen con cierta letra en un minuto. El afásico se desempeña mal en esta tarea, al igual que algunos pacientes con dificultades para buscar y organizar información. Un paciente puede empezar mirando todo el espacio donde se aplica la prueba buscando objetos cuyo nombre empiece con la letra designada. Ésta no es una estrategia productiva. Otros pacientes pueden mostrar perseveración, como cuando dicen lo siguiente si se les pide palabras que empiecen con "A": "albaricoque, arándano, durazno, naranja". En este caso, comenzaron de manera correcta, pero hay perseveración en la categoría de frutas.

Si fue evidente durante la aplicación que un paciente tenía dificultades para comprender las instrucciones orales, podría realizarse una evaluación de los déficits en el lenguaje receptivo. La Prueba de Fichas (*Token Test*) es una medida que se usa con frecuencia para evaluar la comprensión de lenguaje del paciente. Frente al paciente se ponen fichas de distintos tamaños, formas y colores; se dan órdenes con dificultad creciente. Por ejemplo, se le puede pedir que señale el cuadro rojo; después, que señale el cuadro rojo después de tocar el círculo amarillo. Se ha encontrado que esta prueba tiene un muy buen nivel de discriminación entre pacientes afásicos y no afásicos. También existe una versión para niños.

Capacidad visoespacial/perceptual

Las subpruebas de las escalas de inteligencia Wechsler brindan información útil sobre el funcionamiento visoespacial y perceptual. Por ejemplo, Diseño con cubos requiere que el

paciente acomode cubos idénticos con dos lados rojos, dos blancos y con lados mitad rojos mitad blancos, de modo que se vean como los patrones que se le presentan; el diseño puede ser de 2 × 2 o de 3 × 3. Un paciente puede tener un desempeño pobre en esta prueba a causa de una **apraxia construccional**, que consiste en la incapacidad para acomodar o copiar objetos bi- o tridimensionales. El componente manual puede limitarse aplicando la Prueba de organización visual de Hooper, que se compone de 30 reactivos que consisten en un objeto cortado en cuatro partes. La tarea del paciente es acomodar las piezas mentalmente y luego decir el nombre del objeto. Así, se podría encontrar un paciente con un desempeño normal en la Prueba de organización visual de Hooper y un desempeño pobre en Diseño con cubos. Esto reflejaría la dificultad del paciente en la construcción del objeto más que en la percepción. Estos pacientes podrían comentar mientras trabajan en Ensamble de objetos que "saben que es un caballo", pero no pueden "unir las piezas de la manera correcta".

Otra prueba de habilidades visoespaciales, así como de organización, es la Prueba de la figura compleja de Rey/Osterreith. Se pide al paciente que copie un diseño geométrico que tiene muchos detalles; la calificación se determina de acuerdo con la presencia o ausencia de componentes críticos, la exactitud de la ubicación de las características y el grado de distorsión. A menudo, el neuropsicólogo clínico también considera los componentes cualitativos del desempeño del paciente. ¿Trató de hacer el dibujo de una manera lógica? ¿Copió la figura de una manera poco sistemática que resultó en un dibujo distorsionado? ¿Terminó el dibujo con rapidez o lo hizo con mucha lentitud y meticulosidad?

En el caso de pacientes que pueden tener demencia, se puede usar otro tipo de tarea de dibujo. Adivina qué se hace en la Prueba de dibujo de relojes. Correcto. Se da un círculo ya dibujado al paciente o se le pide dibujarlo y luego se le pide que coloque los números en el círculo para formar la carátula de un reloj. A menudo, la instrucción es colocar las manecillas marcando las "once y diez". Como podemos ver en la figura 10-4, algunos pacientes tienen problemas para espaciar los números, otros perseveran y ponen demasiados números, y otros más tienen dificultades para dibujar las manecillas. Algunos incluso colocan los números en el sentido opuesto; a veces, las pruebas visoespaciales descubren la presencia de **desatención espacial**, por lo general en el campo visual izquierdo, resultado de un daño en el hemisferio derecho (figura 10-5).

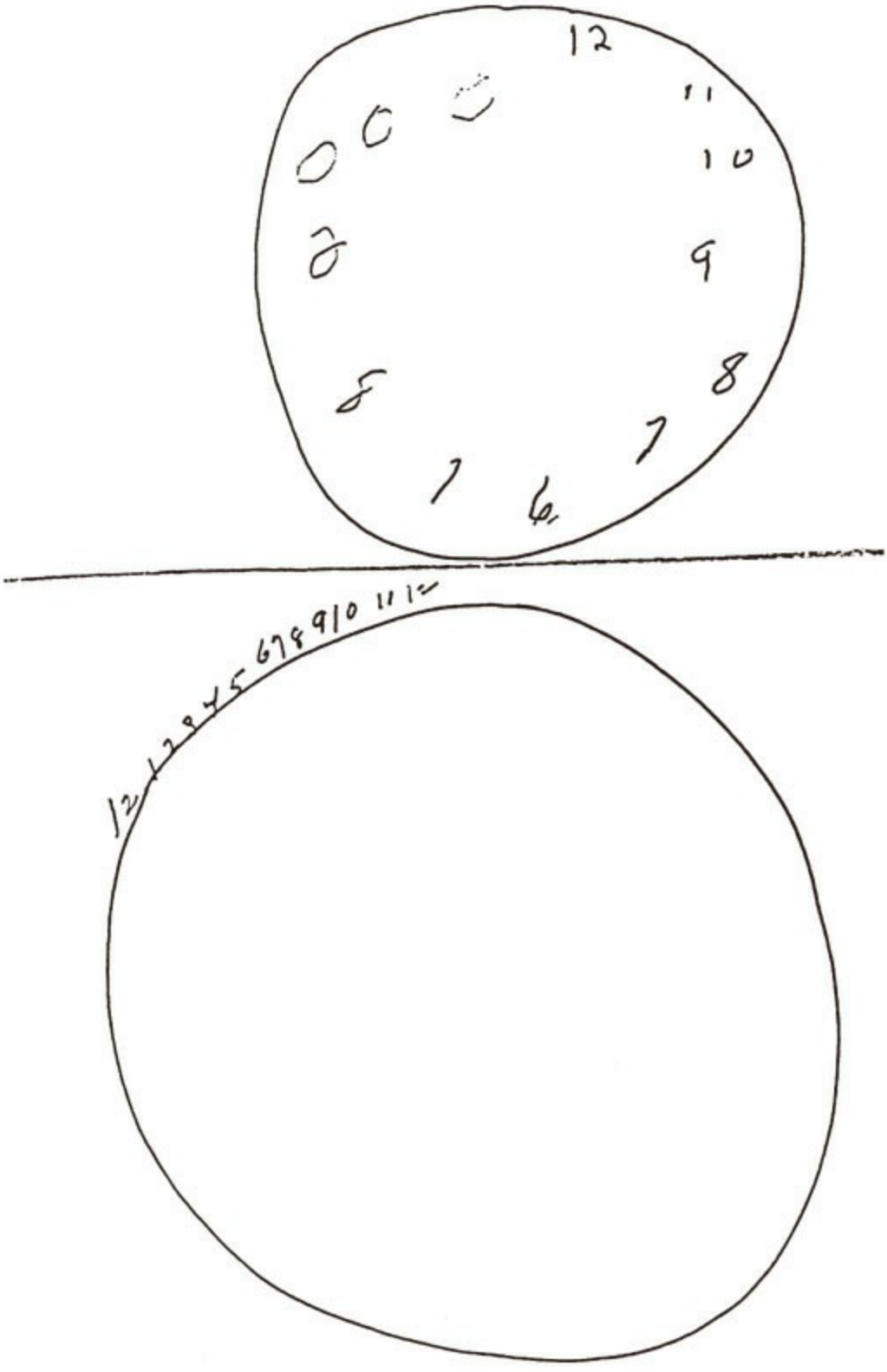


Figura 10-4. Muestras de dibujos de relojes hechos por pacientes con demencia.
Fuente: Cortesía de Brooke J. Cannon, Marywood University.

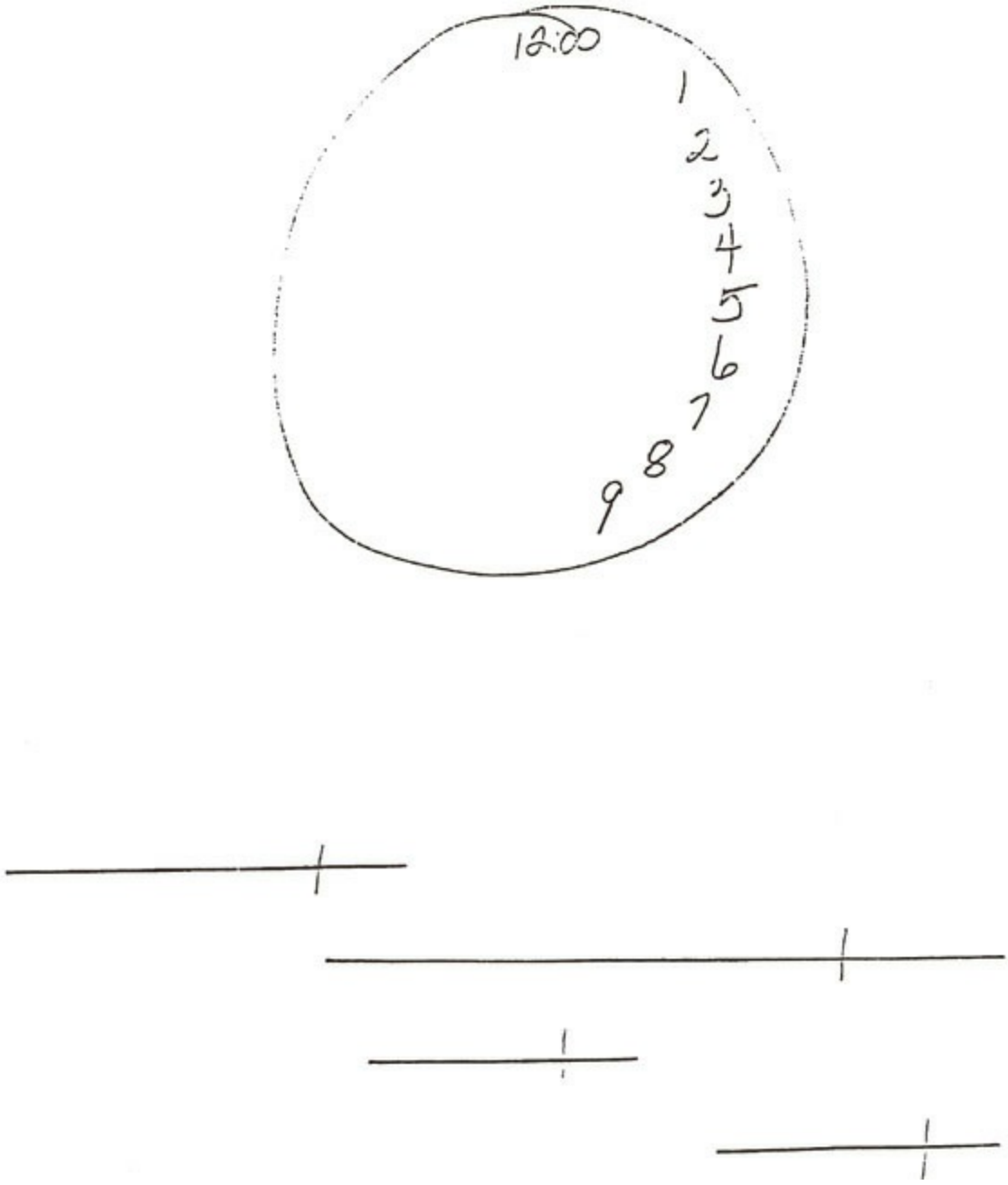


Figura 10-5. Reloj y bisección lineal hechos por un paciente con daño en el hemisferio derecho.
Fuente: Cortesía de Brooke J. Cannon, Marywood University.

¡Inténtalo!

Podemos usar tareas de dibujo para evaluar el funcionamiento cognitivo. Algunos clínicos piden a sus pacientes dibujar una bicicleta. Dibuja una ahora y regresa al texto.

Bien, ¿cómo lo hiciste? Lezak (1995) ideó un sistema de calificación de 20 puntos para los dibujos de bicicletas. La autora considera la presencia de dos llantas (tú probablemente las hiciste), rayos, asiento y manubrio. Además, ¿hay una cadena? ¿Está sujetada de manera apropiada? ¿Tiene velocidades? ¿Y qué hay de las salpicaderas? No te preocupes si te faltaron éstas; la puntuación promedio que se ha informado de obreros de 20 a 24 años de edad es 13.95, con una desviación estándar de 4.03 (Nichols, citado en Lezak, 1995). En la figura 10-6 se puede ver el dibujo de una bicicleta hecho por un paciente con demencia frontotemporal. Observa la llanta delantera.

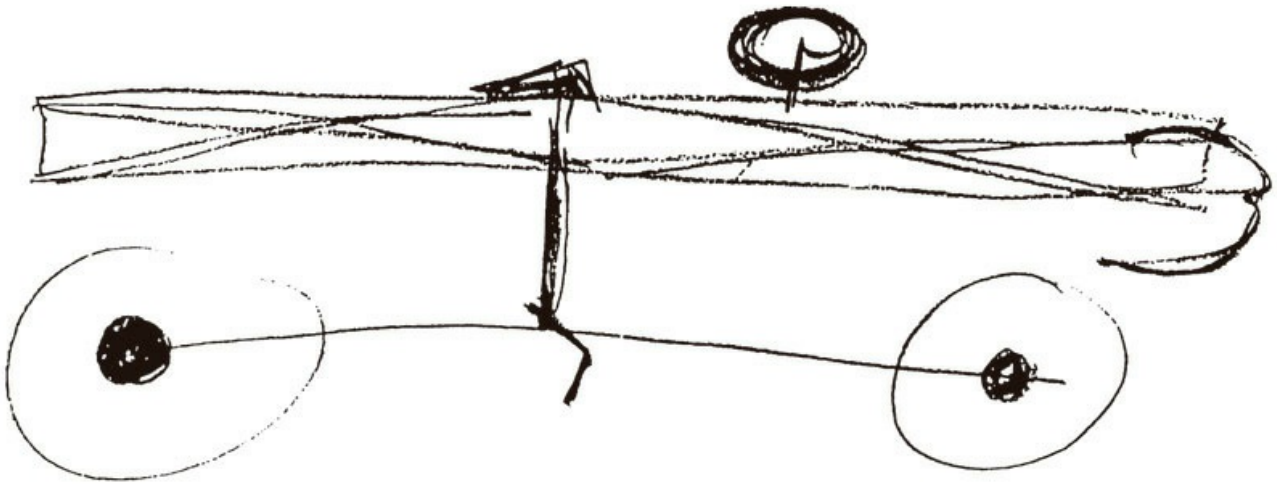


Figura 10-6. Bicicleta dibujada por un paciente con demencia de lóbulo frontal.
Fuente: Cortesía de Brooke J. Cannon, Marywood University.

Memoria

Los déficits de memoria son, quizá, el motivo más frecuente por el que los pacientes son enviados a una evaluación neuropsicológica según lo indican los informes. El deterioro de la memoria puede tener diversas causas, como lesión en la cabeza, derrame cerebral, deficiencia vitamínica, depresión o demencia. Existen pruebas de memoria verbales y no verbales; las verbales pueden implicar recordar una historia corta o pares de palabras, como en la Escala de Memoria de Wechsler, o una larga lista de palabras, como en la Prueba de Aprendizaje Verbal Auditivo de Rey (RAVLT, siglas en inglés). En el RAVLT se presenta una lista de 15 palabras, de la cual el paciente tiene que recordar tantas como pueda y en cualquier orden. La lista se presenta un total de cinco veces y el paciente recuerda todas las palabras que puede después de cada aplicación. Entonces se presenta una lista distractora de 15 nuevas palabras que el paciente tiene que recordar. En este punto se pide al paciente que recuerde las palabras de la primera lista.

Veinte o 30 min más tarde también se pide el recuerdo demorado de la primera lista. La memoria de reconocimiento se evalúa pidiendo al paciente que identifique palabras de la lista original que aparecen entre palabras distractoras. Observando el desempeño del paciente en los cinco primeros ensayos, se puede ver si ha ocurrido o no un aprendizaje.

La mayoría de las personas aprende cerca de cinco palabras del ensayo 1 al 5. A menudo "pierden" dos palabras al intentar recordarlas después de que se presenta la lista distractora. La retención suele ser buena después de la demora, y los pacientes muestran una buena memoria de reconocimiento.

La memoria visual se evalúa con frecuencia pidiendo al paciente que dibuje figuras de memoria. La Escala de Memoria Wechsler-IV contiene varias subpruebas que evalúan la memoria no verbal. Existe una tarea que demanda la reproducción de figuras visuales, otra que evalúa la memoria de varios diseños de 4 × 4 y toma en cuenta exactitud, contenido y ubicación espacial, y una tarea de retención de símbolos en la memoria de trabajo similar al formato de Retención de dígitos que discutimos al hablar del WAIS-IV.

Funcionamiento motor

Los neuropsicólogos clínicos evalúan por lo común tres áreas del funcionamiento motor: velocidad motriz, coordinación motriz fina y fuerza de prensión. La velocidad motriz se puede reflejar de manera indirecta en el desempeño del paciente en otras tareas que requieren de respuestas motrices, como la Prueba de trazo y algunas subpruebas del WAIS. La Prueba de golpeteo dactilar, antes discutida como parte de la Batería Neuropsicológica Halstead-Reitan, se aplica a cada mano; por lo general, la mano dominante tiene un desempeño 10% mejor que la no dominante. La Prueba de tablero de clavijas perforado requiere que el paciente coloque pequeñas clavijas de metal en una serie de agujeros que parecen ojos de cerradura, pero orientadas en distintas direcciones. El tiempo para terminar todo el tablero determina la puntuación, y se presta atención también al número de veces que se le cae una clavija al paciente. La fuerza de prensión, medida con un **dinamómetro** de mano, también se puede evaluar. Otra vez se toman en cuenta las diferencias entre la mano dominante y la no dominante. El dinamómetro de mano también es parte de la Batería Neuropsicológica Halstead-Reitan.

¡Inténtalo!

¿Sabes qué se necesita para ser un ejecutivo exitoso? Enumera las habilidades que piensas que serían importantes. ¿Qué pedirías hacer a un paciente para evaluar estas capacidades?

Funciones ejecutivas

Otras pruebas miden lo que a menudo se denomina **funciones ejecutivas**. El daño cerebral puede alterar

estas habilidades, sobre todo cuando se ubica en los lóbulos frontales.

Las pruebas que requieren **flexibilidad cognitiva**, la capacidad de cambiar de una clase cognitiva a otra, se usan a menudo para medir las "funciones cognitivas". Una medida sensible al deterioro de las funciones cognitivas se basa en el **efecto de Stroop**.

Stroop fue el primero en informarlo en 1935 y se refería, en su origen, a la lentitud para nombrar los colores de tinta cuando el estímulo es incongruente con el nombre del color. La tarea del paciente es nombrar el color de la tinta e ignorar la palabra; ya que leer es una conducta automática cuando se presenta una palabra, nombrar más lentamente el color es resultado de la interferencia en la respuesta. A menudo, primero tenemos que leer la palabra en silencio y luego nombrar el color. La Prueba de colores y palabras de Stroop (Golden, 1978) es una de varias formas de esta prueba. En ésta, hay tres hojas de estímulos, cada una compuesta por cinco columnas de 20 estímulos. La primera hoja contiene simplemente las palabras "rojo", "azul" y "verde" impresas en tinta negra. El paciente tiene que leer tantas palabras como sea posible en 45 segundos. La segunda hoja contiene series de letras XXXX impresas con tinta roja, azul o verde; ahora, el paciente tiene que nombrar el color de la tinta de cada estímulo tantas veces como sea posible en 45 segundos. En la hoja final se encuentran los estímulos de interferencia con cada uno de los nombres de color (rojo, azul, verde) impresos en un color de tinta incongruente. Una vez más, el paciente tiene que nombrar el color de la tinta. El grado de interferencia se determina mediante la diferencia en la velocidad entre nombrar el color de las series XXXX y los nombres de color incongruentes. Se ha encontrado que este efecto de inhibición aumenta en pacientes con daño cerebral; también el estado de ánimo (p. ej., depresión, ansiedad) disminuye el desempeño en esta tarea.

Se han creado algunas variaciones interesantes del efecto de Stroop para usarse con poblaciones clínicas. Por ejemplo, los fóbicos son más lentos para nombrar el color de la tinta de palabras relacionadas con sus temores. Cuando este fenómeno se aplica para propósitos no cognitivos, se denomina "efecto emocional de Stroop" (véase en Williams, Mathews, & MacLeod, 1996, una revisión completa).

Otro ejemplo de una medida de funcionamiento cognitivo es la Prueba de clasificación de tarjetas de Wisconsin. Se presentan al paciente cuatro estímulos que varían entre sí de tres maneras; se da al paciente un conjunto de cartas, cada una de las cuales debe corresponder a uno de los estímulos de alguna manera. Como lo realizó en un ejercicio anterior, primero podría ser por forma, luego por sombreado o por color. Después de colocar cada carta, se dice al paciente si la respuesta es correcta o no; esta información puede usarse entonces para determinar dónde colocar la siguiente carta. Por ejemplo, si la primera respuesta fue de acuerdo con la forma, pero la categoría correcta era el sombreado, el examinador diría "incorrecto". Una vez que el número requerido se haya emparejado de acuerdo con un criterio predeterminado (p. ej., primero por forma), el examinador cambia la regla para que otro criterio sea el correcto. Los pacientes con déficits en las funciones ejecutivas pasan momentos muy difíciles al tener que cambiar de clase cognitiva; un paciente con una demencia que afecta los lóbulos frontales colocó las 128 tarjetas de acuerdo con el mismo criterio a pesar de que se le dijo una y otra vez que eso era "incorrecto". Simplemente no podía cambiar de clase cognitiva y ver los estímulos de una nueva manera. La Prueba de categorías de la Batería Neuropsicológica Halstead-Reitan usa una estrategia similar para ofrecer retroalimentación al paciente con el fin de determinar si está usando una estrategia correcta.

¡Inténtalo!

Divide estos seis estímulos en dos grupos de acuerdo con alguna característica. Ahora divídelos en dos nuevos grupos usando una estrategia diferente. ¿Podrías proponer una tercera manera de separarlos en dos grupos?

Personalidad/estado psicológico

La mayoría de neuropsicólogos clínicos agregan algún tipo de evaluación de la personalidad a su batería de pruebas. Smith, Gorske, Wiggins y Little (2010) encontraron que en la evaluación de problemas de aprendizaje y cuestiones forenses y psiquiátricas es donde se emplean con mayor frecuencia pruebas de personalidad, sobre todo las que son de naturaleza objetiva, que discutiremos en el capítulo 12. Los neuropsicólogos clínicos utilizan muy poco las medidas proyectivas, que revisaremos en el capítulo 13. De acuerdo con una encuesta de Sweet, Moberg y Suchy (2000), 70% de quienes usan pruebas proyectivas lo hace cuando existe la sospecha de un diagnóstico psiquiátrico, pero sólo las utiliza 10% de quienes sospechan o saben de un diagnóstico neurológico.

Entre las medidas objetivas de personalidad, el Inventario Multifásico de Personalidad de Minnesota (MMPI-2) es el que se usa por lo general. El MMPI es el que los neuropsicólogos usan con mayor frecuencia (Camara, Nathan, & Puente, 2000; Smith *et al*., 2010). Describiremos esta prueba con mayor detalle en el capítulo 13. El MMPI-2 se puede aplicar para ayudar a hacer un diagnóstico diferencial (p. ej., trastorno psiquiátrico o trastorno "orgánico"), para detectar posibles cambios de personalidad después de una lesión cerebral (p. ej., paranoia importante), para evaluar la validez del informe de paciente y su motivación (p. ej., configuración de escalas de validez cuestionables) y para determinar el impacto de la disfunción cerebral en el estado psicológico del paciente (p. ej., depresión reactiva).

El MMPI emplea la noción de códigos de dos puntos (véase capítulo 13); algunos de estos códigos se relacionan con lesiones cerebrales. Graham (2006) sugirió que un perfil 1-9 (hipocondriasis-manía) refleja un afrontamiento pobre en una persona que ha sufrido una lesión cerebral. Un perfil 2-9 (depresión-manía) podría reflejar un control emocional deteriorado o hiper-reactividad como un intento de afrontar los resultados de la lesión cerebral. Sin embargo, Graham hizo hincapié en que el MMPI no debe usarse para diagnosticar una lesión cerebral. Ya que el MMPI es una prueba de autoinforme, se da

por hecho que el paciente es capaz de dar un retrato exacto de su estado psicológico. Sin embargo, muchos padecimientos neurológicos imposibilitan la introspección o provocan la negación o inconsciencia de los déficits. Pensemos otra vez en la historia de Phineas Gage, el trabajador de ferrocarril a quien una barra de metal atravesó la cabeza. Si él contestara el MMPI, tal vez negaría cualquier problema a pesar de que sabemos que sus déficits fueron importantes. Sin embargo, sus amigos contarían una historia por completo diferente. Por ello, obtener información de personas cercanas al paciente sobre su funcionamiento puede ser decisivo.

Un tipo de herramienta para la evaluación de la personalidad que incluye los informes de personas cercanas es el *Neuropsychology Behavior and Affect Profile* (NBAP; Nelson, Satz, & D'Elia, 1994), cuya revisión apareció recientemente como NBAP-D (Nelson, Satz, & D'Elia, 2009). Este inventario contiene afirmaciones sencillas acerca de conductas y disposiciones organizadas en cinco escalas: depresión, manía, indiferencia, impropiedad y pragnosia, definida como "un déficit en la pragmática del estilo de comunicación" (Nelson *et al*., 2009, p. 19).

En cada afirmación, el examinado indica si la conducta meta estaba presente *antes* de la lesión o no y si está presente ahora. De esta manera, es posible ver qué áreas han cambiado en relación con los niveles previos a la lesión. El examinado puede hacer una valoración de *sí mismo* o de una *persona cercana*. Así, tenemos la estructura que se bosqueja en el cuadro 10-6: en cada reactivo, el examinado se valora a sí mismo antes y después, o valora a otra persona antes y después. Los reactivos son los mismos, pero redactados de acuerdo con las respuestas: si son sobre uno mismo o sobre otra persona (p. ej., "Parezco deprimido o triste" o "Parece deprimido o triste").

Cuadro 10-6. Estructura del *Neuropsychology Behavior and Affect Profile-D*

Escala	Descripción breve	Reactivos	Formato en primera persona		Formato en tercera persona	
			Antes	Después	Antes	Después
Indiferencia	Minimizar el padecimiento	12				
Impropiedad	Conducta extraña	7				
Pragnosia	Comunicación extraña	12				
Depresión	Estado de ánimo disfórico	11				
Manía	Actividad, estado de ánimo elevado	24				

Cannon (2000) comparó los dos formatos del NBAP y encontró que los pacientes con lesiones cerebrales traumáticas moderadas y las personas cercanas a ellos concuerdan en que los pacientes tienen más problemas en Depresión e Impropiedad después de la lesión. De particular interés fue el hallazgo de que los pacientes reportaron significativamente más problemas con pragnosia después de la lesión en comparación con las personas cercanas a ellos. Esto sugiere que en el caso de las lesiones moderadas, el

paciente puede ser el mejor informante. Estos resultados demuestran la importancia de obtener información de múltiples fuentes.

El estado psicológico puede evaluarse mediante medidas de autoinforme muy usadas como el *Inventario de Depresión de Beck-II* (BDI-II) (véase pp. 347-348a») o mediante una entrevista estructurada diseñada para determinar la presencia de cualquier trastorno psiquiátrico diagnosticable. En una encuesta reciente de los instrumentos de evaluación de la personalidad se encontró que el BDI-II es el que se incluye con mayor frecuencia en las baterías neuropsicológicas, seguido de las escalas de valoración conductual y el MMPI-2 (Smith *et al*., 2010).

Esfuerzo/motivación

Hasta hace poco, muchos neuropsicólogos clínicos se basaban sólo en las escalas de validez del MMPI para establecer la motivación del paciente. Las elevaciones extremas de la escala F en alguien que, por ejemplo, no tiene una psicosis evidente podrían sugerir que el paciente finge estar mal, mientras que las elevaciones en la escala Mentira podría sugerir un intento ingenuo de fingir estar bien. Ya que el MMPI no fue diseñado para evaluar la disfunción neurocognitiva, no ha demostrado ser un método confiable para detectar el fingimiento de síntomas neuropsicológicos. Las escalas F y K son sensibles ante los intentos de fingir algo inusual, principalmente síntomas psicóticos. Si el paciente ha sufrido una lesión craneal moderada, bien podría estar **fingiendo** una enfermedad u otros déficits por una ganancia secundaria, pero el problema que se informa suele ser en la memoria y otras funciones cognitivas, áreas que no incluye el MMPI.

Se han hecho intentos para usar el MMPI-2 con el fin de detectar fingimiento. Se ha desarrollado una "**Escala de fingimiento de mala imagen**" del MMPI-2 para ayudar a detectar la exageración o simulación de tensión emocional en quienes afirman haber sufrido una lesión personal (Lees-Haley, English, & Glenn, 1991). En el capítulo 13 se pueden encontrar las descripciones de las escalas de validez del MMPI-2: L, K, F, TRIN, VRIN, etc. Todas estas escalas pueden formar parte del cuadro cuando se interpreta una evaluación neuropsicológica.

Como se señaló, fingirse enfermo en una evaluación neuropsicológica implica simular síntomas cognitivos, más que una perturbación psiquiátrica. Lo que es aún más notable es que los pacientes fingen deterioro de la memoria. Algunas de las primeras técnicas estandarizadas para medir la veracidad de los pacientes que se quejaban de la memoria empleaban la probabilidad estadística y eran denominadas pruebas de validez de síntomas. Por ejemplo, si se lanzara al aire 100 veces una moneda y alguien adivinara si cae "sol o águila", acertaría 50 veces por puro azar. Utilizando una prueba de elección forzada, por lo general con sólo dos opciones, puede calcularse la probabilidad de respuesta correcta/incorrecta. Así, yo podría mostrar una pluma roja o azul a un paciente y pedirle que recuerde el color; podría esperar 10 segundos y luego preguntarle el color de la pluma que le acabo de mostrar. Ahora, si hago esto 100 veces (en realidad, esto suele hacerse con una computadora), incluso si el paciente tiene un deterioro grave de

memoria, debe acertar al menos 50% de las ocasiones. Ahora imaginemos que el paciente está fingiendo déficits de memoria; al no tener un deterioro real, el paciente recordaría con facilidad la respuesta correcta, pero entonces nos diría intencionalmente la respuesta equivocada. Así, si el paciente acierta en menos de 50% de las ocasiones (de hecho, fuera del intervalo de confianza calculado), es probable que de manera deliberada diga las respuestas equivocadas, pues sólo así el paciente podría equivocarse más de lo que predice el azar.

Otras pruebas también examinan la exactitud en reactivos fáciles y difíciles, así como la latencia de la respuesta, que a menudo se mide pidiendo que el paciente responda en el teclado de una computadora. Debería tomar más tiempo identificar la respuesta correcta y luego elegir la incorrecta que responder honestamente. Sin embargo, este fenómeno puede ocultarse si quien se finge enfermo simula un funcionamiento motor y cognitivo más lento.

Ciertas pruebas, como el *Test of Memory Malingering* (TOMM), han desarrollado puntuaciones de corte para determinar si el examinado no se desempeña al nivel de sus capacidades. En el TOMM se muestra al paciente una serie de 50 dibujos sencillos; después se le muestran conjuntos de dos dibujos y se le pide que elija cuál dibujo se le había mostrado antes. Tombaugh (1996) demostró que, en esta tarea, incluso los individuos con demencia son capaces de obtener al menos 45 reactivos correctos de 50 posibles, y que no muestran pérdida significativa de la memoria en un periodo de demora de 15 min. Es decir, el TOMM fue diseñado para ser tan fácil que sólo los individuos que conscientemente tratan de decir respuestas equivocadas olvidarían más de 10% de los reactivos. El desarrollo de medidas como el TOMM ayudará en el futuro a los neuropsicólogos a detectar el fingimiento.

¡Inténtalo!

Imagina que estás fingiendo un daño cerebral debido a un accidente automovilístico. Estás demandando al otro conductor y te han enviado a una evaluación neuropsicológica. ¿Cómo te desempeñarías en las diferentes medidas que hemos descrito? ¿Piensas que podrías fingir una lesión creíble? ¿Hay áreas cognitivas en las que te desempeñarías con normalidad?

Información complementaria

Una evaluación neuropsicológica completa implica más que aplicar e interpretar pruebas. Aquí describimos, de manera muy breve, otros tipos de información que se examina, por lo general, al realizar una evaluación neuropsicológica.

Historia médica

Es decisivo obtener la historia médica completa del paciente al llevar a cabo una evaluación neuropsicológica. Cuando el neuropsicólogo clínico trabaja en un escenario de hospitalización, estos datos suelen estar disponibles; en el expediente médico se encuentra una historia médica realizada por el médico tratante. Aquí, el neuropsicólogo clínico se entera de cualquier posible contribución médica al estado cognitivo y psicológico del paciente. Por ejemplo, una disfunción tiroidea puede relacionarse con un estado de ánimo deprimido, o un déficit vitamínico puede parecer una demencia. Una historia de enfermedad cardiovascular, como ritmo cardiaco irregular, cirugías del corazón o ataques cardiacos, puede poner al paciente en riesgo de un derrame cerebral. Los déficits visuales pueden dificultar la capacidad del paciente para ver los estímulos de la prueba. Una historia de enfermedad neurológica o lesión cerebral pasada también constituyen información crítica. La historia familiar de enfermedades también es importante, en especial la de trastornos que puedan ser heredados, como la enfermedad de Huntington.

Leer todas las notas del progreso que hace el equipo de tratamiento también puede ser muy instructivo. Se puede saber si ha habido conductas inusuales, si el paciente ha mostrado signos de psicosis o paranoia o si ha habido períodos de confusión durante el día o la noche. En algunos casos de demencia se puede presentar un estado de confusión: el paciente está razonablemente bien orientado durante el día, pero se siente muy confundido durante la noche. El expediente médico también permite al neuropsicólogo clínico determinar qué medicamentos está tomando el paciente. Idealmente, los informes de TAC o IRM del cerebro están disponibles. Además, el pabellón de trabajo social a menudo realiza una historia psicosocial detallada. Así, revisar el expediente antes de ver al paciente permite obtener mucha información y formular hipótesis.

Muchos pacientes no se atienden en el hospital y, por lo tanto, es necesario obtener esta misma información del paciente externo. Si el paciente no se ha practicado recientemente un examen físico, debería hacerlo antes de comenzar con la evaluación neuropsicológica. Se puede pedir el expediente al médico tratante, y al paciente o a un familiar suyo, que llene formatos con los antecedentes del paciente; la información que se obtenga puede detallarse o clarificarse en una entrevista previa a la evaluación.

Historia psiquiátrica

El neuropsicólogo clínico debe indagar acerca de tratamientos pasados o actuales de padecimientos psiquiátricos. ¿Alguna vez ha estado hospitalizado el paciente por razones psiquiátricas? ¿Alguna vez ha estado en psicoterapia o en tratamiento psiquiátrico? Cuando se presenta un deterioro de la memoria, también es pertinente saber si en la historia del paciente ha habido terapia electroconvulsiva (TEC) o "tratamiento de choques eléctricos".

Algunos pacientes pueden informar una historia de perturbación psicológica, pero nunca han recibido tratamiento. Por medio de una entrevista detallada, el neuropsicólogo clínico podría intentar determinar diagnósticos posibles y la gravedad de estas dificultades. Otra vez, la historia familiar psiquiátrica constituye información importante que es necesario tener.

Historia psicosocial

Es importante conocer los antecedentes educativos del paciente. ¿Hasta dónde llegó en la escuela? ¿Qué notas obtenía? ¿Recibió algún tipo de educación especial? ¿Alguna vez fue retenido, suspendido, expulsado? ¿Cuáles eran sus materias favoritas?

El neuropsicólogo clínico preguntaría acerca de la familia de origen del paciente y estado actual de los padres y hermanos. También se debe obtener información sobre la historia ocupacional del paciente, estado marital y número de hijos. Además, es necesario determinar si hubo o hay uso de drogas y alcohol; también sería pertinente cualquier problema legal. Asimismo son importantes las actividades de esparcimiento o el trabajo como voluntario, pues pueden ayudar a comprender el estado actual y pasado; en particular, es importante señalar los cambios en el nivel de actividad.

Registros escolares

En particular en el caso de los pacientes más jóvenes, los registros escolares ofrecen una gran riqueza de información. A menudo se han realizado pruebas de inteligencia, hay puntuaciones de pruebas nacionales de aprovechamiento disponibles y el maestro comenta sobre la conducta del paciente; todo esto puede ser muy valioso. Las notas aportan evidencia de las fortalezas y debilidades; siempre es interesante comparar el autoinforme del paciente ("Fui un estudiante de 9") con su expediente real. Muy pocas veces afirman haber sido peores de lo que realmente fueron; de hecho, suele ser al contrario. Así, los registros escolares pueden brindar información acerca del estado premórbido, historia de educación especial y validez del autoinforme del paciente.

Información colateral

Como se señaló al discutir el NBAP, reunir información directamente del paciente, así

como de otra fuente, puede ser muy útil para crear un cuadro completo del estado del paciente. Sobre todo en casos en que el paciente ya no puede hablar de su funcionamiento, los miembros de la familia proporcionan información decisiva. Consideremos un caso de demencia temprana en el que la persona niega tener cualquier dificultad con la memoria u otros problemas. Era un hombre agradable y su comportamiento social era apropiado, aparentaba ser una fina persona. Sin embargo, cuando se entrevistó a su esposa y después durante la evaluación, se hicieron evidentes los déficits en áreas importantes.

Observaciones conductuales

Otro componente importante de la evaluación neuropsicológica es la observación de la conducta del paciente. Por lo general, los neuropsicólogos clínicos incluyen esta observación en sus informes escritos, pues pueden ilustrar el tipo de errores que comete el paciente. Por ejemplo, la manera en que un paciente aborda la prueba puede decir más que el hecho de responder correcta o incorrectamente un reactivo. Podría señalarse que cuando se le pidió copiar un diseño muy complejo, no empleó una estrategia sino que lo hizo de una manera poco sistemática. O cuando realiza la subprueba Diseño con cubos de las escalas de inteligencia Wechsler, el paciente no logró mantener la configuración de 2 × 2 o 3 × 3. A veces, el paciente amontonaba en efecto los cubos para formar una figura tridimensional.

Si el paciente no se muestra cooperativo con la aplicación o parece distraído e incapaz de enfocarse en la tarea, esta información debe incluirse en el informe. Cualquier otra conducta inusual que se observe durante la evaluación puede ser importante en las consideraciones diagnósticas, como la evidencia de las actividades de esparcimiento, deterioro auditivo o visual o déficits en el habla espontánea.

Resumen de puntos clave 10-3

Información complementaria reunida en una evaluación neuropsicológica
Historia médica
Historia psiquiátrica
Historia psicosocial
Registros escolares
Información colateral
Observaciones conductuales

De vuelta a los casos

Ahora retomaremos los casos descritos al inicio del capítulo; en cada caso, veremos cómo ciertas pruebas neuropsicológicas o patrones de resultados ayudarían a responder las preguntas relacionadas con el motivo de consulta. El hecho de que sólo ciertas pruebas se presentan como evidencia no significa que sean las únicas medidas aplicadas; el espacio no nos permite discutir toda la información que se reuniría o presentar todos los resultados de pruebas.

Caso #1:

> Nancy O'Roarke, mujer de 74 años de edad que recientemente enviudó, se queja de pérdida de memoria, problemas para concentrarse e incapacidad para realizar sus actividades cotidianas. ¿Tiene la enfermedad de Alzheimer?

Imaginemos que la señora O'Roarke tiene buena salud: no hay historia de problemas médicos importantes o uso de drogas o alcohol, ni perturbaciones o tratamientos psiquiátricos previos.

Los resultados de un examen médico reciente fueron normales. Su infancia fue normal, sin problemas médicos ni psiquiátricos en la familia. En el bachillerato fue la estudiante con las calificaciones más altas y tuvo altas notas en la universidad. La señora O'Roarke es una maestra de primaria jubilada. Tiene una hija de 45 años que trabaja como contadora y vive con su esposo cerca de su madre. La paciente estuvo casada 48 años, hasta que, hace dos meses, su esposo murió de manera repentina a causa de un ataque cardiaco. En las últimas cuatro semanas se ha quejado de ser olvidadiza, tener problemas para concentrarse y ser incapaz de continuar con su trabajo de voluntaria y hacer sus labores domésticas.

Así, ¿cuáles son las posibilidades? ¿Podría tratarse de una demencia temprana? No es inusual que los pacientes sean enviados a una evaluación neuropsicológica en busca de demencia después de la pérdida del cónyuge. En muchos casos, el paciente ha tenido problemas cognitivos que pasaron desapercibidos por un largo periodo, pues el cónyuge se encargaba de todas las responsabilidades. Por lo que al morir el cónyuge y dejar solo al paciente, se vuelve evidente que éste no puede vivir de manera independiente.

En la evaluación neuropsicológica, el desempeño de la señora O'Roarke fue inconsistente. Su memoria verbal, cuando tuvo que recordar una historia, tuvo un deterioro moderado. En contraste, cuando se le pidió que aprendiera una larga lista de palabras, mostró una curva de aprendizaje de promedio a superior al promedio. Sin embargo, su recuerdo demorado de estas palabras presentó un deterioro moderado, pero tuvo una memoria de reconocimiento normal. La memoria visual fue de promedio a superior en tareas de recuerdo inmediato y demorado. La atención mostró deterioros en la mayoría de las medidas. Las habilidades de lenguaje fueron de nivel promedio a

superior y las habilidades visoespaciales se ubicaron dentro del promedio, aunque en las tareas cronometradas fue un poco lenta. En las pruebas motrices su desempeño fue normal. Cuando se le aplicó la prueba de inteligencia Wechsler, la señora O'Roarke obtuvo un índice de Comprensión verbal de 117, promedio alto, y un índice de Razonamiento perceptual de 108, promedio. Se observó que no logró obtener puntos de bonificación por terminar rápido las tareas de ejecución.

Así, ¿qué piensas de los resultados de las pruebas? ¿Existe un deterioro de la memoria? ¿Y qué hay de los déficits de atención y la motricidad lenta? ¿Se trata de demencia?

La señora O'Roarke respondió el MMPI-2. Su perfil mostró una tendencia a negar problemas, pues todas las escalas clínicas cayeron dentro del rango normal con excepción de Depresión e Introversión social, las cuales alcanzaron el rango de clínicamente significativas.

La hija de la paciente informó que su madre había sido una mujer activa y social hasta la muerte de su padre. La señora O'Roarke se ocupó de los arreglos del funeral y del papeleo financiero. Una vez que esto terminó, simplemente dejó de comprometerse en actividades fuera de casa. No comía de manera apropiada y empezó a bajar de peso, además de que se quejaba de tener problemas para dormir por las noches. Cuando su hija hablaba con ella, la paciente tenía dificultades para prestarle atención mientras conversaban, pues era incapaz de concentrarse.

Ahora, ¿qué piensas? Sin duda, es probable la presencia de depresión de acuerdo con el MMPI-2, el informe de la hija, la reciente muerte del esposo y la presencia de síntomas comunes en la depresión de ancianos, como dificultades para dormir y pérdida de apetito, de peso y de interés en actividades usuales.

Sin embargo, ¿la depresión explica los déficits cognitivos? En muchos casos, la respuesta es "¡Sí!" Con base en estos datos, sería inadecuado diagnosticar demencia; más bien, parece tratarse de una **seudodemencia**, que es un deterioro cognitivo parecido a la demencia de tipo Alzheimer, pero que se relaciona con un padecimiento psiquiátrico, por lo general, depresión. Cuando se sospecha que un paciente tiene seudodemencia, debe enviarse a un tratamiento para la depresión que incluya medicamentos antidepresivos y/o psicoterapia. Cuando se resuelve la depresión, los déficits cognitivos suelen desaparecer.

Cuadro 10-7. Características de la evaluación neuropsicológica de la seudodemencia

Respuestas "no sé" frecuentes
Tendencia a rendirse con facilidad, pero continúan si se les anima
Errores por omisión en vez de por comisión (es decir, no dan respuestas, mientras que en la demencia dan respuestas incorrectas)
La memoria de reconocimiento es mejor que la de recuperación
Atención/concentración deteriorada
Desorientación
No hay deterioro de lenguaje
Estado de ánimo deprimido

La señora O'Roarke muestra un patrón típico de resultados de pruebas relacionados con seudodemencia (cuadro 10-7), pues tuvo un desempeño inconsistente en tareas de memoria, a veces normal y a veces con deterioro. Si se tratara de demencia, en todas las pruebas aparecería un deterioro sin excepción. Además, la señora O'Roarke mostró una diferencia significativa en su recuerdo espontáneo de información, que estaba deteriorado, en comparación con su memoria de reconocimiento, que fue normal. Esto es como la diferencia entre preguntas tipo "llena los espacios en blanco" y las de opción múltiple; si no sabemos la respuesta, lo haremos igual de mal en cualquiera de las dos. Lo mismo sucede con el paciente con demencia: tiene dificultades para almacenar información, por lo que no hay nada que pueda ser recuperado. Si **sí** sabemos la respuesta pero no la podemos recordar en cierto momento, no podremos llenar el espacio en blanco, pero quizá podamos contestar correctamente si la respuesta es de opción múltiple, pues podemos reconocer la respuesta correcta entre las opciones. Esto es como el paciente anciano deprimido que almacenó de manera adecuada la información, pero tiene problemas para recuperarla.

Caso #2:

Jack Davis, trabajador de la construcción de 42 años de edad, sufrió una lesión en su trabajo cuando un bloque de cemento lo golpeó en la cabeza. Desde entonces, tiene dolores de cabeza, pérdida de memoria, períodos breves de atención y un estado de ánimo deprimido. Se encuentra en un pleito legal contra su patrón. ¿Tiene daño cerebral?

En el momento de la lesión, el señor Davis traía un casco de seguridad cuando el bloque de cemento cayó de una altura de 7 m, lo golpeó en el casco y lo derribó al suelo; después, lo llevaron a la sala de urgencias de un hospital local. En ningún momento perdió la conciencia. El examen médico reveló una IRM normal del cerebro y ningún déficit neurológico. El señor Davis había sido razonablemente saludable hasta el accidente. Fuma media cajetilla de cigarros al día y bebe cerveza en reuniones sociales. No hay antecedentes de perturbaciones o tratamientos psiquiátricos.

A pesar de no haber signos de daño cerebral de acuerdo con la IRM, ¿podría haber sufrido una lesión que explicara sus déficits? Veamos más información.

El señor Davis informa haber tenido una infancia normal, sin problemas médicos ni psiquiátricos en su familia. Su padres están vivos y bien de salud, y sus dos hermanas están casadas y tienen hijos. El señor Davis era un atleta en la escuela; en la universidad, practicó lucha libre, en la que fue campeón estatal, y futbol. Cuando se le preguntó por sus calificaciones, el paciente afirmó haber sido un estudiante "promedio"; negó cualquier problema importante en la escuela, pero reconoció que sólo disfrutaba los deportes.

Actualmente, el paciente está en proceso de divorcio debido a "diferencias irreconciliables"; ha estado separado de su esposa desde el año pasado. Tiene dos hijos de 12 y 15 años. Informa que el menor de ellos fue diagnosticado con trastorno por

déficit de atención y toma Ritalín. Niega haber tenido problemas legales antes de esta demanda.

Cuando se le preguntó por el accidente, el señor Davis describió con un tono enojado cómo le había dicho al encargado de la obra que no derribara el muro sin la malla de contención, pero "¡él no me hizo caso!" El señor Davis describió cómo su amigo lo había llevado al hospital donde una enfermera grosera lo examinó y lo "tuvieron esperando 90 min" antes de ver a un doctor. El paciente también enumeró los distintos exámenes que le han hecho desde el accidente, algunos ordenados por su abogado y otros por el de la compañía constructora. No ha regresado al trabajo desde el accidente y demanda la compensación correspondiente.

El señor Davis no dio su consentimiento para contactar a su esposa o a su patrón con el fin de obtener información adicional, pero sí autorizó que se consultara sus expedientes escolares, en los que se encontró que varias veces se fue de pinta y una vez lo suspendieron por pelearse. Sus calificaciones más comunes eran 7 y 8, pero 10 en mecánica automotriz, taller de carpintería y educación física. La evaluación de su inteligencia en la escuela reveló un CI Total de 95, pero las pruebas de aprovechamiento indicaban un rendimiento cada vez menor de lo que se esperaba de acuerdo con su nivel de inteligencia.

Los resultados de las pruebas indicaron, en general, un desempeño de promedio a y promedio bajo en las subpruebas de inteligencia, con excepción de Retención de dígitos, Aritmética y Símbolos y dígitos, en los que se encontró un deterioro moderado. Su capacidad visoespacial fue normal y la de nombrar objetos fue promedio, pero su fluidez verbal (medida mediante la generación de listas de palabras) mostró un deterioro grave. La memoria verbal y visual también mostró un deterioro grave tanto en el recuerdo inmediato como en el demorado. No mostró una curva de aprendizaje en el *Key Auditory Verbal Learning Test*, y su memoria de reconocimiento fue pobre, pues cometió varios errores en la identificación de estímulos. Su motricidad fue promedio alto en ambas manos. En la prueba de aprovechamiento se encontró un deterioro significativo en las habilidades de deletreo y lectura, pues su desempeño corresponde al nivel de quinto grado; en aritmética utilizando lápiz y papel tuvo un nivel promedio bajo.

¿Qué piensas de los resultados hasta aquí? ¿Son consistentes con lo que imaginabas que podría ser causado por una lesión como ésta? Sin duda, las lesiones craneales cerradas producen déficits de atención/concentración y deterioro de la memoria.

En el MMPI-2, el señor Davis aprobó reactivos que indican aflicción psicológica; la escala F fue alta y la K, mucho más baja. Todas las escalas clínicas, excepto Masculinidad/Feminidad (las puntuaciones bajas indican intereses masculinos estereotípicos) e Introversión social, fueron significativamente altas. Entonces se aplicó el Inventario de Depresión de Beck-II, cuyos resultados sugieren un nivel moderado de depresión.

Se llevó a cabo la Prueba de validez de síntomas con un procedimiento de elección forzada entre dos opciones. El señor Davis tuvo correcto 40% de los ensayos. Se quejó con frecuencia de su incapacidad para recordar incluso las cosas más sencillas.

Ahora, con estos datos, podríamos suponer que el señor Davis está fingiendo un deterioro de la memoria. A pesar de que sus puntuaciones en las pruebas de memoria caen en el rango de deterioro, fue capaz de proporcionar información muy precisa y detallada acerca de su accidente y los subsiguientes tratamientos médicos. Los síntomas que declaró en el MMPI-2 son exagerados, pues están en un nivel que indicaría psicosis, lo cual indica que finge estar enfermo o que se trata de "un grito de ayuda". El esfuerzo/motivación durante la evaluación sugiere que proporcionó de manera deliberada respuestas incorrectas.

Los expedientes escolares sugieren que el señor Davis no fue bueno para la escuela; de hecho, quizá aprobó los grados debido a sus méritos atléticos. Las pruebas de aprovechamiento, tanto las escolares como las de la evaluación actual, pueden indicar un problema de aprendizaje. Su hijo tiene diagnóstico de trastorno por déficit de atención, el cual suele presentarse en varios miembros de las familias. Por lo tanto, pueden esperarse algunos deterioros que no están relacionados con el accidente. Sin embargo, parece que el señor Davis exagera sus déficits.

Así, ¿el señor Davis sólo está siendo codicioso y trata de obtener una indemnización monetaria a pesar de que no tiene una lesión real? Es posible. También debemos considerar su situación actual. Está en un amargo proceso de divorcio, no está trabajando y no tiene ingresos. De acuerdo con el Inventario de Depresión de Beck-II, puede estar experimentando niveles moderados de depresión (esta escala no contiene una indicación de la validez de las respuestas del paciente).

Por lo tanto, la información no permite apoyar de manera concluyente la existencia de una lesión cerebral producida por el accidente. Las pruebas de validez de los síntomas sugieren un fingimiento intencional o una exageración de los déficits. La imposibilidad de obtener información colateral, sin duda, hace más difícil determinar los cambios posteriores al accidente. También es posible que esté sufriendo una depresión situacional y que tenga un problema de aprendizaje premórbido.

El último caso nos permitirá explorar más la evaluación de los problemas de aprendizaje.

Caso #3:

> Billy Zollinger, niño de 10 años de edad de cuarto grado, ha bajado sus notas en la escuela. Su madre informa que a menudo pierde cosas, olvida hacer lo que ella le pide o no logra terminar sus trabajos. Además, le toma un tiempo excesivo hacer sus tareas escolares, y a menudo se distrae con ruidos exteriores y otros estímulos. ¿Tienes trastornos por déficit de atención?

La historia médica de Billy incluye complicaciones en el parto. Fue un "bebé triste", sufrió de hipoxia, una disminución de oxígeno, y tuvo que estar internado en una unidad de cuidados intensivos para recién nacidos, donde fue entubado. Ahí permaneció durante una semana. Después de ser dado de alta y durante los años preescolares, no hubo retrasos en el desarrollo, es decir, habló, caminó, etc., de acuerdo con lo esperable.

Padeció de "chichones en la cabeza", según su madre, casi siempre debidos a su gusto por subirse a los árboles y a su activa conducta durante juegos en que se ponía en riesgo, como colgarse de cabeza. Sufrió una fractura en la muñeca al caerse en el patio de recreo a los 7 años. Aunque disfruta las actividades físicas, su madre lo describe como un niño "con poca coordinación".

Otras pruebas médicas indican visión y audición normales; no hay déficits vitamínicos, sus resultados de laboratorio, peso y estatura son normales para su edad. Su lateralidad predominante es la izquierda, pero no hay otros miembros de la familia inmediata ni extensa con la misma lateralidad que él.

El maestro de Billy informa que es un niño agradable, pero que tiene dificultades para seguir las instrucciones que da en clase y a menudo necesita que se las repita de manera individual. Con frecuencia no puede terminar su trabajo en el tiempo asignado. En su trabajo hay errores menores, como sumar en lugar de restar.

Los resultados de las pruebas neuropsicológicas reflejan capacidades intelectuales de promedio a promedio alto, pero también déficits significativos en las subpruebas relacionadas con la atención del WISC-IV, como Aritmética, Retención de dígitos, Búsqueda de símbolos y Figuras incompletas. Su aprovechamiento académico está por debajo del nivel de su grado en deletreo y matemáticas. Sus errores escritos en matemáticas a menudo son equivocaciones simples. La Batería de Pruebas Neuropsicológicas Halstead-Reitan mostró déficits consistentes con los problemas de atención auditiva, dificultad con la percepción de sonidos del habla y deterioro relativo en la mano derecha en comparación con la izquierda. En la evaluación del lenguaje escrito se encontró que Billy es bastante descuidado al escribir a mano, y hubo errores de deletreo y omisión de algunas palabras. La evaluación de la lectura reveló que el funcionamiento, en general, está intacto, pero se presentan errores sencillos, como leer "carro" en vez "corro". La evaluación de la personalidad y el estado de ánimo no indicó anormalidades.

Esta breve descripción de la evaluación de Billy pone de relieve los resultados que a menudo se encuentran en un niño con trastorno por déficit de atención, tipo inatentivo. Todas las pruebas con un fuerte componente de atención/concentración reflejan un deterioro en relación con estímulos visuales y auditivos. Las pruebas de lenguaje reflejan la presencia de dislexia, un problema de aprendizaje caracterizado por un deletreo pobre y errores de lectura. El patrón del niño en su desempeño en las pruebas determina el tipo específico de problema de aprendizaje que puede estar presente (cuadro 10-8).

Cuadro 10-8. Definiciones de tres problemas de aprendizaje específicos

Problema de aprendizaje	Definición
Dislexia diseidética	Incapacidad para leer palabras como un todo, de modo que el niño debe pronunciar en voz alta toda la palabra. Los errores de lectura ocurren en palabras como "caballo/cabaña". Los errores en la escritura son fonéticos, como "taco" en vez de "Paco".

Dislexia disfonética	Incapacidad para pronunciar en voz alta palabras, así que la lectura es de palabras completas y depende del vocabulario visual. Los errores de deletreo no son fonéticos, como "instuito" en vez de "instituto". Los errores de lectura son sustituciones de palabras visualmente similares, como "donito" en vez de "bonito".
Discalculia	Las capacidades matemáticas son inferiores a lo esperado para la edad, educación e inteligencia. Los déficits pueden incluir dificultades para comprender conceptos matemáticos, reconocer símbolos numéricos, llevar apropiadamente los números en una cuenta y realizar operaciones aritméticas.

La definición original de un problema de aprendizaje requería una discrepancia significativa entre el nivel de inteligencia del niño y su aprovechamiento. Una revisión reciente de la ley federal de EUA en relación con la educación de estudiantes con estos problemas ahora se enfoca más en el fracaso para responder a los métodos educativos tradicionales (Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C.; IDEA, 2004). En el capítulo 16 se encuentra una discusión sobre IDEA 2004.

El patrón del desempeño en la evaluación determina el tipo específico de problema de aprendizaje que el niño puede presentar (cuadro 10-8). Las tasas de prevalencia de los problemas de aprendizaje se estiman en 2-10%, con 5% de los niños de escuelas públicas con diagnóstico de problemas de aprendizaje (American Psychiatric Association, 2000).

Resumen

1. No siempre se pensó que el cerebro se encargara de controlar la conducta. Gracias a las contribuciones de Alcmeón, Galeno, Gall, Broca y Wernicke, se reveló el papel crítico del cerebro.
2. El campo de la neuropsicología clínica se desarrolló durante el siglo XX, y continúa haciéndolo en el siglo XXI. La profesión ha alcanzado su apogeo en 50 años.
3. Las razones para hacer una evaluación neuropsicológica incluyen diagnóstico, evaluación de fortalezas y debilidades cognitivas, ayuda al tratamiento o la planeación vocacional, uso en contextos forenses e investigación.
4. La evaluación neuropsicológica puede llevarse a cabo siguiendo el método de batería fija o el de batería flexible. La Batería Neuropsicológica Luria-Nebraska y la Batería Neuropsicológica Halstead-Reitan son ejemplos de baterías fijas. La batería flexible permite que el clínico elija las pruebas que, considera, son las mejores para evaluar al cliente.
5. Los dominios cognitivos que se pueden evaluar durante una valoración neuropsicológica incluyen atención, aprovechamiento, esfuerzo/motivación, funciones ejecutivas, inteligencia, lenguaje, memoria, estado mental, habilidades motrices, personalidad/estado psicológico y capacidad visoespacial/perceptual.
6. La historia médica, los expedientes escolares, la historia psiquiátrica, la información colateral, la historia psicosocial y las observaciones conductuales del paciente son importantes para tener un cuadro neuropsicológico completo.
7. La seudodemencia se puede presentar de cierta manera como una verdadera demencia, pero los resultados de la evaluación neuropsicológica tienen un patrón diferente.
8. Algunos pacientes pueden fingir déficits cognitivos, en especial por una ganancia secundaria. Sin embargo, existen métodos de evaluación neuropsicológica que pueden ayudar a detectar a quien finge estar enfermo.
9. Dislexia diseidética, dislexia disfonética y discalculia son tres ejemplos de problemas de aprendizaje frecuentes.

Palabras clave

afasia
apraxia construccional
batería fija
batería flexible
Batería Neuropsicológica Halstead-Reitan
Batería Neuropsicológica Luria-Nebraska
Broca, Paul
desatención espacial
dinamómetro
discalculia
dislexia diseidética
dislexia disfonética
efecto de Stroop
Escala de fingimiento de mala imagen
fingirse enfermo
flexibilidad cognitiva
frenología
funciones ejecutivas
Galeno
Gall, Franz Josef
Índice de deterioro
neuropsicología
neuropsicología clínica
perseverar
premórbido
Puntuación de Déficit Neuropsicológico General
seudodemencia
Wernicke, Carl

Ejercicios

1. Lee los textos originales de *Paul Broca* en Classics in the History of Psychology, sitio de internet diseñado por Christopher D. Green de la Universidad de York, Toronto, Ontario, en la dirección http://psychclassics.yorku.ca/
2. Explora la presencia de la neuropsicología clínica en internet. Un buen sitio para comenzar es el de Neuropsychology Central en http://www.neuropsychologycentral.com/. En un buscador, introduce el término neuropsicología y revisa los resultados.
3. Conoce más del desafortunado trabajador de ferrocarril, *Phineas Gage*. Visita su página en http://www.uakron.edu/gage/
4. Usa las siguientes subpruebas del WAIS-IV para determinar las fortalezas y debilidades del paciente. Todas éstas son puntuaciones corregidas por edad. Calcula el promedio de las subpruebas. Encierra en un círculo las que están tres puntos por encima o por debajo del promedio y escribe F (fortaleza) o D (debilidad) junto a ellas.

Diseño con cubos = 10 Semejanzas = 12 Retención de dígitos = 7 Matrices = 8 Vocabulario = 14	Aritmética = 7 Búsqueda de símbolos = 12 Rompecabezas visual = 11 Información = 13 Claves = 6
Promedio =	

5. Trata de medir el *control mental*. Cuenta hacia atrás de 3 en 3 empezando de 100. Deletrea la palabra "chocolate" en orden inverso. Recita los meses del año en orden inverso empezando con diciembre.
6. Prueba tu *fluidez verbal*. Ve cuántas palabras puedes escribir en 60 segundos que empiecen con la letra M, después L y después D. Ahora intenta con la *fluidez de categorías*. Anota tipos de vegetales, luego animales y luego colores. ¿Qué es más fácil, letras o categorías? ¿Por qué?
7. Entra al sitio de internet del *Mental Measurement Yearbook* de Buros en http://buros.org/. Introduce como palabras clave cada uno de los *dominios cognitivos* evaluados con una batería flexible (véase cuadro 10-4). Lee más acerca de las pruebas y determina si medirían de manera apropiada esa capacidad cognitiva.
8. Prueba tu *memoria*. Nombra los últimos cinco presidentes de México en orden inverso, empezando con el presidente actual. Nombra los 32 estados. Dibuja un billete sin verlo.
9. Con ayuda de las pruebas que discutimos para cada dominio cognitivo (cuadro 10-4), predice cómo sería el desempeño de una persona con *trastorno por déficit de atención*. ¿En qué pruebas su desempeño sería el peor? ¿Qué pruebas no serían afectadas?

10. Crea preguntas que podrían evaluar el *juicio* o la *solución de problemas* de una persona, es decir, ambos tipos de *funciones ejecutivas*. Por ejemplo, "¿Qué harías si no hubiera agua en tu casa y necesitaras lavarte los dientes? Ve qué soluciones pueden dar tus amigos a estos problemas.

CAPÍTULO 11
Pruebas de aprovechamiento

Objetivos

1. Usar el continuo capacidad-aprovechamiento para distinguir entre las pruebas que se basan en mayor o menor entrenamiento específico.
2. Esbozar un programa de evaluación escolar típico.
3. Definir los movimientos de responsabilidad y estándar.
4. Enumerar las principales categorías de pruebas de aprovechamiento[1].
5. De cada categoría, dar uno o dos ejemplos de pruebas reales.
6. Enumerar los usos típicos de las pruebas de cada categoría.
7. Identificar las características de las pruebas de cada categoría.

Introducción

Las pruebas de aprovechamiento comprenden una vasta y diversa serie de pruebas. En términos de la cantidad absoluta de pruebas, las de aprovechamiento superan de manera abrumadora a todos los otros tipos de pruebas combinadas. Esto es particularmente cierto si incluimos el casi infinito número de pruebas "hechas por el maestro" y las evaluaciones permeadas por los libros de texto que se aplican a diario en todos los niveles educativos y en cada rincón del planeta. Pero esta generalización es cierta incluso si excluimos las pruebas hechas por el maestro y nos limitamos a contar las pruebas más formales y estandarizadas. Para ayudar a organizar esta vasta colección de material, será útil empezar nuestra cobertura de las pruebas de aprovechamiento con una orientación y un panorama general. Primero, distinguimos entre pruebas de capacidad y de aprovechamiento; luego, consideramos el papel de los psicólogos en el campo de las pruebas de aprovechamiento. Seguimos con un esquema de clasificación de este tipo de pruebas, el cual utilizamos para organizar el resto del capítulo. Por último, ofrecemos un contexto propio para algunas pruebas que consideramos y describimos un programa típico de evaluación escolar y el movimiento de responsabilidad en la educación.

Continuo capacidad-aprovechamiento

Suele hacerse una distinción entre las pruebas de capacidad y las de aprovechamiento; las primeras las revisamos en capítulos previos. Aunque la separación de pruebas de capacidad y de aprovechamiento en capítulos diferentes es habitual y práctica, lo mejor es pensar en estas pruebas como parte de un continuo más que compartimientos con una clara separación. Este continuo se presenta de manera gráfica en la figura 11-1. El **continuo capacidad-aprovechamiento** representa el grado en que el entrenamiento específico influye en el desempeño en la prueba. En el extremo derecho se encuentran las pruebas que dependen en gran medida del entrenamiento específico, por ejemplo, conocimiento del programa Excel, batallas de la Guerra Civil, procedimientos de contabilidad, reglamentos locales o la capacidad para andar en bicicleta, jugar beisbol o navegar en bote. En el extremo izquierdo se encuentran las capacidades que, se piensa, son sumamente generales, por ejemplo, detectar patrones numéricos, crear anagramas o simplemente conocer el significado de muchas palabras. Estas últimas capacidades, sin duda, se desarrollan como resultado de algunas experiencias, pero no son tan específicas.

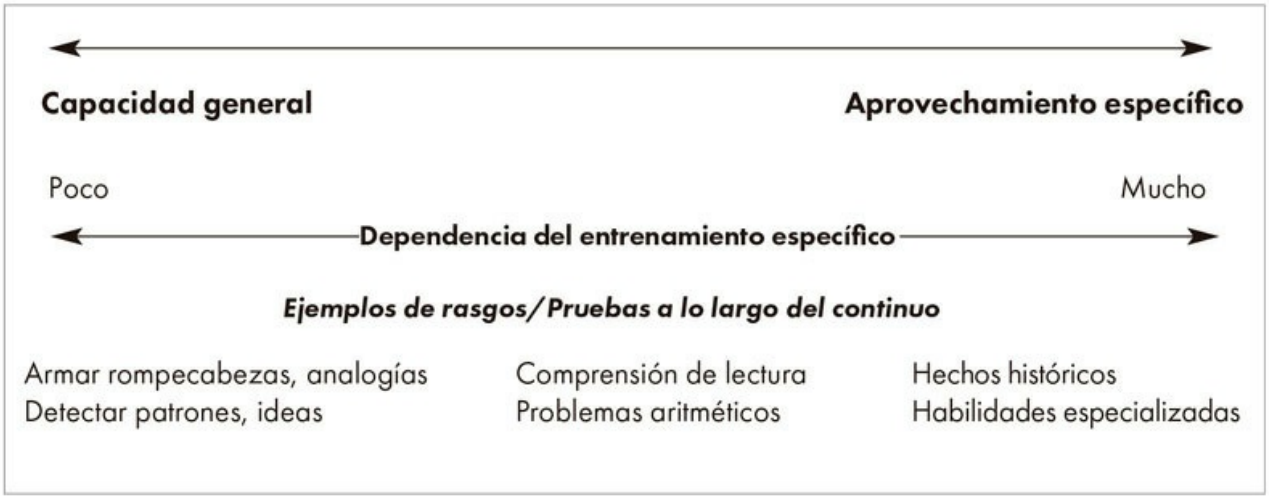


Figura 11-1. Continuo capacidad-aprovechamiento.

Observa que algunas capacidades caen en la mitad del continuo. La solución de problemas aritméticos depende, en parte, del entrenamiento específico con los hechos numéricos que se usan en el problema y, en parte, de la capacidad más general de analizar lo que debe hacerse con determinados números. La comprensión de lectura depende, en parte, de las habilidades para analizar palabras y, en parte, de la capacidad para descifrar las ideas más importantes de un pasaje. Será útil tener en mente este continuo cuando consideremos al amplio rango de pruebas de aprovechamiento tratadas en este capítulo.

Relación del psicólogo con las pruebas de aprovechamiento

Las pruebas de capacidad mental, en especial las de aplicación individual, y las de personalidad, tanto objetivas como proyectivas, constituyen el repertorio del psicólogo. Las pruebas de aprovechamiento se identifican tradicionalmente más en el campo de la educación que en el de la psicología. Sin embargo, los psicólogos se relacionan con el campo de las pruebas de aprovechamiento de muchas maneras importantes. Revisemos brevemente éstas antes de comenzar el tratamiento formal de esta categoría de pruebas.

Primero, a causa de su formación especial en psicometría, los psicólogos, con frecuencia, tienen un papel importante en el desarrollo de las pruebas de aprovechamiento, lo cual requiere conocimiento profundo del área de contenido que cubre la prueba, así como de los principios de la construcción de pruebas. Los educadores u otros profesionales suelen aportar el conocimiento del contenido y los psicólogos, el conocimiento de la construcción de pruebas.

Segundo, varios subcampos de la psicología tienen sus principales aplicaciones en escenarios escolares, como la psicología escolar y la orientación. Las personas de estas profesiones con frecuencia trabajan con los resultados de pruebas de aprovechamiento y,

por lo general, se centran en los individuos. Sin embargo, debido a su formación en la metodología de evaluación e interpretación, los psicólogos escolares y orientadores a menudo forman parte de comités de pruebas que eligen las pruebas de aprovechamiento, hacen informes de los resultados distritales a los consejos escolares y cumplen con otros deberes.

Tercero, muchos psicólogos sin relación directa con las escuelas reciben con frecuencia informes de los resultados de estas pruebas; por ejemplo, un psicólogo clínico infantil a menudo obtiene el expediente escolar del niño, que contiene resultados de pruebas estandarizadas de aprovechamiento, entre otras cosas. Su trabajo es integrar toda la información que obtenga sobre el niño, lo cual es de especial importancia al tratarse de casos de problemas de aprendizaje o trastorno por déficit de atención con hiperactividad (TDAH).

Por último, las pruebas de aprovechamiento tienen un papel importante en la investigación en muchos campos de la psicología. Un psicólogo del desarrollo puede usar estas pruebas, junto con otras medidas de capacidad mental, para estudiar el impacto de las experiencias infantiles tempranas. Un neuropsicólogo puede estudiar la relación entre ciertas funciones cerebrales y el patrón de resultados en una batería de aprovechamiento. Un psicólogo social puede estudiar los resúmenes de estas pruebas en relación con las características socioeconómicas de naciones enteras.

En resumen, al mismo tiempo que las pruebas de aprovechamiento tienen un papel importante en el campo de la educación, también son importantes en psicología. El psicólogo necesita conocer la elaboración de estas pruebas, su estructura típica y los tipos de puntuaciones que se obtienen de ellas.

Clasificación amplia de las pruebas de aprovechamiento

Para los fines de este capítulo, clasificamos las pruebas de aprovechamiento en seis categorías amplias, como se muestra en el cuadro 11-1, lo cual se debe más a la comodidad práctica que a la elegancia teórica. En la práctica, los límites entre estas categorías son permeables. La primera categoría incluye baterías de aprovechamiento que se usan mucho en los programas de evaluación de escuelas primarias y secundarias. La segunda contiene pruebas de aprovechamiento de área única que se utilizan primordialmente en programas educativos de bachillerato y superiores, así como en contextos del ámbito laboral. La tercera incluye pruebas de aprovechamiento hechas por encargo de programas de evaluación estatales, nacionales e internacionales. La cuarta categoría contiene numerosos exámenes de licencia y certificación que usan las organizaciones profesionales. La quinta incluye pruebas de aprovechamiento de aplicación individual que se usan, por lo general, junto con pruebas de capacidad mental para diagnosticar problemas de los estudiantes; a veces se les denomina baterías psicoeducativas. Por cada tipo de pruebas, identificamos ejemplos, usos típicos y características comunes.

Cuadro 11-1. Clasificación amplia de las pruebas de aprovechamiento

Baterías de pruebas de aprovechamiento	Pruebas de aprovechamiento de área única
Exámenes de licencia y certificación	Pruebas estatales, nacionales, internacionales
Baterías psicoeducativas	Pruebas hechas por el maestro

No tratamos la amplia categoría de las pruebas hechas por los maestros, ni las que se elaboran para el salón de clases, ni las que se usan en los programas de formación en la industria o el ejército. Si este libro estuviera dirigido principalmente a educadores, dedicaríamos más espacio a explorar las directrices para elaborar y aplicar pruebas para el salón de clases, los procedimientos de calificación y temas relacionados. Sin embargo, es una decisión pensada no abordar estos temas, que se pueden consultar en Hogan (2007), Miller, Linn y Gronlund (2009) y Nitko y Brookhart (2011).

Programa de evaluación escolar típico

Muchas de las pruebas que presentamos en este capítulo aparecen en los programas de evaluación escolar; para poner las pruebas en contexto, será de utilidad bosquejar un programa típico (véase cuadro 11-2). Empezando con los requerimientos de *No Child Left Behind Act* (NCLB), adoptado como ley federal en EUA en 2002, la parte más importante de los programas de evaluación escolar se convirtió en pruebas financiadas por el estado, al menos en las escuelas públicas. Una escuela típica también aplica una batería de aprovechamiento, nuestra primera categoría, en varios grados. Las escuelas privadas, que no emplean las pruebas financiadas por el estado, usan estas baterías de aprovechamiento en más grados; en algunos de ellos, la escuela también utiliza pruebas de capacidad mental de aplicación grupal, como las que describimos en el capítulo 9, aunque esto ha disminuido en años recientes. El programa de evaluación escolar, por lo general, incluye una prueba de intereses vocacionales en los años de bachillerato; estas pruebas se consideran en el capítulo 14. Aunque no son parte formal de estos programas, muchos estudiantes presentan pruebas de admisión a la universidad en el segundo y tercer año del bachillerato. Por último, subgrupos de estudiantes presentan muchas pruebas adicionales que aplican, por ejemplo, los psicólogos escolares u otros especialistas relacionados con programas correctivos, evaluaciones de problemas de aprendizaje, selección para programas de sobredotación intelectual, entre otros.

Cuadro 11-2. Programa de evaluación escolar típico

Tipo de prueba/Grado	**K**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Programa de evaluación estatal				x	x	x	x	x	x			x	
Batería de aprovechamiento			x							x			
Prueba grupal de capacidad				x					x				
Inventario de intereses vocacionales											x		
Pruebas de admisión a la universidad												x	x

Otras pruebas – Casos selectos	x	x	x	x	x	x	x	x	x	x	x	x	x

Movimiento de responsabilidad y educación basada en estándares

El panorama de las pruebas de aprovechamiento no estaría completo sin una mención de los movimientos de responsabilidad y la educación basada en estándares. Estos movimientos han tenido una influencia profunda en cómo se elaboran, usan e interpretan las pruebas de aprovechamiento. En los círculos educativos, la **responsabilidad** se refiere al hecho de que las escuelas son responsables de su producto, que es el aprendizaje del alumno, el cual se refleja a menudo, pero no exclusivamente, en su desempeño en las pruebas de aprovechamiento. De ahí que estas pruebas sean un elemento importante para cualquier discusión de la responsabilidad. Cuando un legislador o un editor de periódico demanda mayor responsabilidad, suele referirse, en términos operacionales, a poner a prueba y ser más críticos con los resultados de las pruebas.

El origen del movimiento de responsabilidad educativa en EUA puede rastrearse más o menos en la década de 1960. Varios eventos o tendencias crearon las bases de este movimiento, de los cuales identificaremos tres factores. Primero, a finales de la década de 1950, EUA estaba impresionado con el Sputnik: el primer vuelo espacial de Rusia (entonces Unión de Repúblicas Socialistas Soviéticas, URSS), que tomó por sorpresa a la comunicad científica y dejó pasmado al público de EUA. En un santiamén, la idea de que ciencia, tecnología y educación de dicho país eran las mejores del mundo quedó destruida y el consiguiente clamor exigiendo una reforma educativa era ensordecedor. En el ambiente dominado por las dos superpotencias políticas de esos días, se trataba literalmente de una cuestión de vida o muerte. El segundo factor fue el aumento radical en los fondos para la educación. Los dólares (ajustados de acuerdo con la inflación) por estudiante en la educación primaria, secundaria y de bachillerato casi se triplicaron de principios de la década de 1950 a principios de la de 1970; en *National Center for Education Statistics* (2001) se puede consultar las cifras exactas. La gente –contribuyentes y legisladores– quería saber qué recibían a cambio de su dinero. Tercero, en 1964 el Congreso de EUA aprobó la ley *Elementary and Secondary Education Act* [Ley de Educación Elemental y Secundaria] (**ESEA**) mediante la cual se proporcionaban fondos federales en una escala sumamente ampliada para una gran cantidad de esfuerzos educativos. Muchas partes de esta legislación estipularon que las iniciativas educativas necesitaban ser evaluadas. Los educadores que comenzaban programas de acuerdo con ESEA tenían que ser responsables de su éxito. La evaluación, por lo general, implicaba pruebas estandarizadas. La ley *No Child Left Behind Act* [Que ningún niño se quede atrás], quizá la fuerza más importante en las escuelas públicas de la primera década del siglo XXI, es técnicamente una revisión de ESEA. Otra ley federal, *Individuals with Disabilities in Education Act of 2004* [Ley de Individuos con Discapacidades en la Educación] (véase capítulo 16), también tiene implicaciones considerables para el uso de pruebas en escenarios educativos.

El movimiento de responsabilidad, que se originó en la década de 1960, evolucionó hasta dar por resultado lo que llamamos **educación basada en estándares**, enfoque que demanda una clara identificación del contenido que deben aprender los alumnos, la especificación de niveles requeridos de desempeño y la certeza de que los estudiantes tienen la oportunidad de aprender el material. Ahora, todos los estados tienen alguna versión del enfoque basado en estándares de la educación. Las pruebas de aprovechamiento se usan para determinar si los estándares se han cumplido. Cizek (2001b) ofrece un excelente resumen del desarrollo del movimiento de los estándares, y Thurlow e Ysseldyke (2001) rastrean su base legislativa para el movimiento.

En el panorama educativo actual, las pruebas estandarizadas de aprovechamiento tienen un papel importante; en algunos estados, los estudiantes necesitan alcanzar cierta puntuación en una prueba creada por el estado o publicada de manera comercial para recibir un diploma de bachillerato o ser admitidos en el siguiente grado. En algunas universidades, los estudiantes tienen que demostrar eficiencia en una prueba antes de continuar con cursos superiores. La entrada en ciertas profesiones depende del desempeño exitoso en pruebas incluso después de haber concluido un programa educativo prescrito. Todos estos son ejemplos de **pruebas de gran importancia**: sus resultados tienen consecuencias inmediatas y considerables para el individuo. Los resultados de las pruebas de toda una escuela o un distrito también pueden tener implicaciones importantes para los administradores escolares. En este caso, de manera interesante, la prueba puede no ser de gran importancia para las personas que la responden (los estudiantes), pero lo es para alguien más (el director de la escuela o el superintendente).

El movimiento de responsabilidad es una fuerza poderosa en la educación moderna de EUA. El público quiere pruebas de que los estudiantes aprenden. Las pruebas de aprovechamiento de un tipo u otro casi siempre forman parte de este cuadro.

Baterías de aprovechamiento «291a

Consideraremos primero las baterías estandarizadas de aprovechamiento que se usan en las escuelas primarias y secundarias. El término **batería** en este contexto significa una serie coordinada de pruebas que cubren diferentes áreas de contenido y niveles de grados múltiples. Existen innumerables ejemplos de varios tipos de pruebas incluidos en este libro, pero esto no ocurre con las baterías estandarizadas de aprovechamiento. Sólo existen cuatro baterías importantes de aprovechamiento en EUA (véase cuadro 11-3). No hemos intentado hacer una lista de baterías usadas en otros países.
«291-300a

Cuadro 11-3. Baterías importantes de aprovechamiento

Batería	Editorial	Página de internet
Iowa Tests of Basic Skills[a]	*Riverside Publishing*	riverpub.com
Metropolitan Achievement Tests	*Pearson Assessment*	pearsonassessment.com
Stanford Achievement Test	*Pearson Assessment*	pearsonassessment.com
TerraNova[b]	*CTB/McGraw-Hill*[c]	ctb.com

[a] Nuestra referencia al *Iowa Tests of Basic Skills* (ITBS) tiene la intención de incluir el *Iowa Tests of Educational Development* diseñado para los grados de bachillerato y en una escala que es continuación de la del ITBS. La editorial se refiere cada vez más a las dos series de pruebas en conjunto simplemente como *los Iowa Tests*.
[b] La edición actual del *TerraNova* ocupa el lugar de ediciones previas que se basaron en dos baterías anteriores: *California Achievement Test* (CAT) y el *Comprehensive Test of Basic Skills* (CTBS).
[c] "CTB" es el acrónimo de California Test Burrau, título retirado de manera oficial, pero que se usa aún en muchas fuentes.

¡Inténtalo!

Visita la página de internet de algunas de las editoriales que aparecen en el cuadro 11-3. ¿Qué información se destaca acerca de la batería de aprovechamiento?

Prueba de Aprovechamiento de Stanford

Ilustramos las baterías estandarizadas de aprovechamiento con la *Prueba de Aprovechamiento de Stanford*, 10a. Edición (SAT10, siglas en inglés de *Stanford Achievement Test*). En general, las otras baterías que se enumeran en el cuadro 11-3 muestran las principales características del SAT10 que se describen más adelante. Al igual que todas estas baterías, el SAT10 es, en realidad, un vasto sistema de medidas más que una prueba única, como será evidente en nuestra presentación. También ocurre que

las ediciones nuevas de estas baterías tienden a aparecer cada cinco o seis años. El SAT10 comprende dos niveles del *Stanford Early School Achievement Test* [Prueba de Aprovechamiento Escolar Temprano de Stanford] (SESAT 1 y 2), las principales series Stanford y tres niveles del *Test of Academic Skills* [Prueba de Habilidades Académicas] (TASK 1, 2 y 3). En el cuadro 11-4 se bosquejan las pruebas y niveles que se incluyen en este sistema.

La revisión del cuadro 11-4 revela varias características del SAT10 que son comunes en las principales baterías estandarizadas de aprovechamiento. Podemos notar, primero, que hay diferentes niveles de la prueba diseñados para distintos grados; por ejemplo, el Nivel P2 (Primaria 2) está diseñado para los grados 2.5-3.5, es decir, desde la mitad del segundo grado hasta la mitad del tercero.

Cuadro 11-4. Stanford 10: alcance y secuencia de opción múltiple

Niveles de la prueba, rangos de grados recomendados, pruebas y tiempos de aplicación																								
Niveles de la prueba	**S1 Grado K.0-K.5 K T**		**S2 Grado K.5-1.5 K T**		**P1 Grado 1.5-2.5 K T**		**P2 Grado 2.5-3.5 K T**		**P3 Grado 3.5-4.5 K T**		**11 Grado 4.5-5.5 K T**		**12 Grado 5.5-6.5 K T**		**13 Grado 6.5-7.5 K T**		**A1 Grado 7.5-8.5 K T**		**A2 Grado 8.5-9.9 K T**		**T1 Grado 9.0-9.9 K T**		**T2 Grado 10.0-10.9 K T**		**T3 Gr 11 12 K**
Sonidos y letras	40	30	40	25																				
Habilidades para estudiar palabras					30	20	30	20	30	20	30	20												
Lectura de palabras	30	15	30	25	30	25																		
Lectura de oraciones			30	30	30	30																		
Vocabulario de lectura							30	20	30	20	30	20	30	20	30	20	30	20	30	20	30	20	30	20	30
Comprensión de lectura					40	40	40	40	54	50	54	50	54	50	54	50	54	50	54	50	54	40	54	40	54
Lectura total	70	45	100	80	130	115	100	80	114	90	114	90	84	70	84	70	84	70	84	70	84	60	84	60	84
Matemáticas	40	30	40	30																	50	50	50	50	50
Resolución de problemas matemáticos					42	50	44	50	46	50	48	50	48	50	48	50	48	50	48	50					
Procedimientos matemáticos					30	30	30	30	30	30	32	30	32	30	32	30	32	30	32	30					
Matemáticas total					72	80	74	80	76	80	80	80	80	80	80	80	80	80	80	80					
Lenguaje					40	40	48	45	48	45	48	45	48	45	48	45	48	45	48	45	48	40	48	40	48
Deletreo					36	30	36	30	38	35	40	35	40	35	40	35	40	35	40	35	40	30	40	30	40
Escucha de palabras e historias	40	30	40	30																					
Habilidades de escucha					40	30	40	30	40	30	40	30	40	30	40	30	40	30	40	30					
Ambiente	40	30	40	30	40	30	40	30																	

Ciencia									40	25	40	25	40	25	40	25	40	25	40	25	40	25	40	25	40	25	40	
Ciencia social									40	25	40	25	40	25	40	25	40	25	40	25	40	25	40	25	40	25	40	
Batería básica	150	105	180	140	318	295	298	265	316	280	322	280	292	260	292	260	292	260	292	260	222	180	222	180	22			
Batería completa	190	135	220	170	358	325	338	295	396	330	402	330	372	310	372	310	372	310	372	310	302	230	302	230	30			
Tiempo total de aplicación	2 hrs. 15 min		2 hrs. 50 min		5 hrs. 25 min.		4 hrs. 55 min.		5 hrs. 30 min.		5 hrs. 30 min.		5 hrs. 10 min.		5 hrs. 10 min.		5 hrs. 10 min.		5 hrs. 10 min.		3 hrs. 50 min.		3 hrs. 50 min.		3 h 50			
Forma de lenguaje D					40	40	40	40	45	45	48	45	48	45	48	45	48	45	48	45	48	40	48	40	48			

K = No. de reactivos

T = Tiempo en minutos

(Estandarizada sin cronometrar; los tiempos dados son con propósitos de planeación únicamente.)

Fuente: *Stanford Achievement Test*, Tenth Edition, Technical Manual. Copyright © 2003 por Harcourt Assessment, Inc. Reproducido autorización. Todos los derechos reservados.

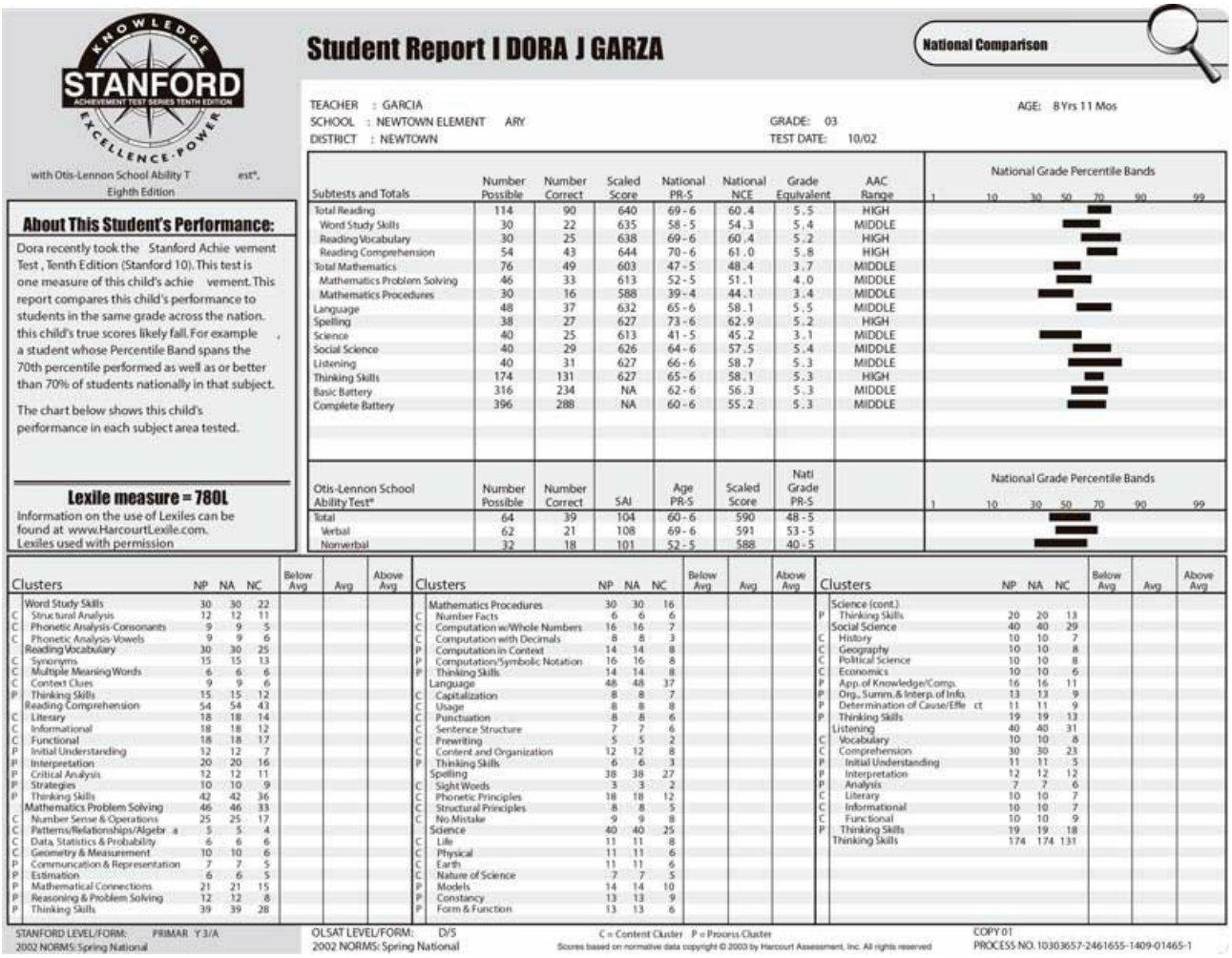
Algunas pruebas de capacidad mental de aplicación grupal, que consideramos en el capítulo 9, también emplean niveles múltiples; cada uno contiene una gran cantidad de pruebas específicas. Esta característica es la que da origen al término batería. Las subpruebas específicas van y vienen en los distintos niveles (moviéndose de izquierda a derecha en el cuadro). Por ejemplo, Sonidos y letras aparece en los dos niveles inferiores y nada más. Las pruebas separadas en Ciencia y Estudios sociales no aparecen sino en el nivel 3 de primaria. A pesar de este fenómeno de ir y venir, hay una gran continuidad a lo largo de las series; por ejemplo, algunas medidas de lectura y matemáticas ocurren en todos los niveles. Todos los niveles también tienen puntuaciones de Batería básica y Batería completa. Una subprueba típica contiene cerca de 40 reactivos y su aplicación requiere cerca de 25 min. Las subpruebas se conjuntan en los totales de área (p. ej., lectura o matemáticas), que suelen tener entre 75 y 100 reactivos. Podemos notar que la proporción entre el número de reactivos y el número de minutos para su aplicación tiende a estar en un rango de 2:1 a 1:1.

Aunque no se señala en el cuadro 11-4, el SAT10 tiene una Batería completa y una Batería abreviada, la cual tiene menos subpruebas que son, en general, más cortas que las de la Batería completa. También hay una prueba de escritura que requiere crear una historia o un ensayo que se califica de acuerdo con el método holístico o analítico.

El SAT10 ofrece casi todo tipo de puntuación derivada, que vimos en el capítulo 3, e incluye rangos percentiles, estaninas,, puntuaciones escalares, equivalentes de grado y equivalentes de curva normal. Los rangos percentiles y las estaninas, se proporcionan tanto en el caso de las puntuaciones individuales como en el de los promedios grupales. También hay comparaciones entre capacidad y aprovechamiento, categorías de desempeño (por debajo del promedio, promedio, por encima del promedio) en los grupos de reactivos relacionados y valores p para los reactivos individuales. Por último, hay "estándares de desempeño", juicios con referencia al criterio acerca de lo adecuado del desempeño en la prueba. Las categorías que se incluyen en los informes son Avanzado, Eficiente, Básico y Por debajo de lo básico.

El SAT10 incluye una gran variedad de informes de calificaciones generados por

computadora. La figura 11-2 muestra uno de los informes estándar populares de un estudiante individual. Podemos notar que el informe proporciona varios tipos de calificaciones normativas, presenta gráficas que incorporan la noción de banda percentil (véase pp. 96a») e incluye los resultados de una prueba grupal de capacidad (*Otis-Lennon School Ability Test* [Prueba de Capacidad Escolar Otis-Lennon]), junto con una comparación capacidad-aprovechamiento.



Student Report I DORA J GARZA

National Comparison

TEACHER : GARCIA
SCHOOL : NEWTOWN ELEMENT ARY
DISTRICT : NEWTOWN
GRADE: 03
TEST DATE: 10/02
AGE: 8 Yrs 11 Mos

with Otis-Lennon School Ability T est®, Eighth Edition

About This Student's Performance:

Dora recently took the Stanford Achie vement Test , Tenth Edition (Stanford 10). This test is one measure of this child's achie vement. This report compares this child's performance to students in the same grade across the nation. this child's true scores likely fall. For example , a student whose Percentile Band spans the 70th percentile performed as well as or better than 70% of students nationally in that subject.

The chart below shows this child's performance in each subject area tested.

Lexile measure = 780L

Information on the use of Lexiles can be found at www.HarcourtLexile.com. Lexiles used with permission

Subtests and Totals	Number Possible	Number Correct	Scaled Score	National PR-S	National NCE	Grade Equivalent	AAC Range
Total Reading	114	90	640	69-6	60.4	5.5	HIGH
Word Study Skills	30	22	635	58-5	54.3	5.4	MIDDLE
Reading Vocabulary	30	25	638	69-6	60.4	5.2	HIGH
Reading Comprehension	54	43	644	70-6	61.0	5.8	HIGH
Total Mathematics	76	49	603	47-5	48.4	3.7	MIDDLE
Mathematics Problem Solving	46	33	613	52-5	51.1	4.0	MIDDLE
Mathematics Procedures	30	16	588	39-4	44.1	3.4	MIDDLE
Language	48	37	632	65-6	58.1	5.5	MIDDLE
Spelling	38	27	627	73-6	62.9	5.2	HIGH
Science	40	25	613	41-5	45.2	3.1	MIDDLE
Social Science	40	29	626	64-6	57.5	5.4	MIDDLE
Listening	40	31	627	66-6	58.7	5.3	MIDDLE
Thinking Skills	174	131	627	65-6	58.1	5.3	HIGH
Basic Battery	316	234	NA	62-6	56.3	5.3	MIDDLE
Complete Battery	396	288	NA	60-6	55.2	5.3	MIDDLE

National Grade Percentile Bands: 1 10 30 50 70 90 99

Otis-Lennon School Ability Test®	Number Possible	Number Correct	SAI	Age PR-S	Scaled Score	Nati Grade PR-S
Total	64	39	104	60-6	590	48-5
Verbal	62	21	108	69-6	591	53-5
Nonverbal	32	18	101	52-5	588	40-5

National Grade Percentile Bands: 1 10 30 50 70 90 99

Clusters	NP	NA	NC	Below Avg	Avg	Above Avg
Word Study Skills	30	30	22			
C Structural Analysis	12	12	11			
C Phonetic Analysis-Consonants	9	9	5			
C Phonetic Analysis-Vowels	9	9	6			
Reading Vocabulary	30	30	25			
C Synonyms	15	15	13			
C Multiple Meaning Words	6	6	6			
C Context Clues	9	9	6			
P Thinking Skills	15	15	12			
Reading Comprehension	54	54	43			
C Literary	18	18	14			
C Informational	18	18	12			
C Functional	18	18	17			
P Initial Understanding	12	12	7			
P Interpretation	20	20	16			
P Critical Analysis	12	12	11			
P Strategies	10	10	9			
P Thinking Skills	42	42	36			
Mathematics Problem Solving	46	46	33			
C Number Sense & Operations	25	25	17			
C Patterns/Relationships/Algebr a	5	5	4			
C Data, Statistics & Probability	6	6	6			
C Geometry & Measurement	10	10	6			
P Communcation & Representation	7	7	5			
P Estimation	6	6	5			
P Mathematical Connections	21	21	15			
P Reasoning & Problem Solving	12	12	8			
P Thinking Skills	39	39	28			

Clusters	NP	NA	NC	Below Avg	Avg	Above Avg
Mathematics Procedures	30	30	16			
C Number Facts	6	6	6			
C Computation w/Whole Numbers	16	16	7			
C Computation with Decimals	8	8	3			
P Computation in Context	14	14	8			
P Computation/Symbolic Notation	16	16	8			
P Thinking Skills	14	14	8			
Language	48	48	37			
C Capitalization	8	8	7			
C Usage	8	8	8			
C Punctuation	8	8	6			
C Sentence Structure	7	7	6			
C Prewriting	5	5	2			
C Content and Organization	12	12	8			
P Thinking Skills	6	6	3			
Spelling	38	38	27			
C Sight Words	3	3	2			
C Phonetic Principles	18	18	12			
C Structural Principles	8	8	5			
C No Mistake	9	9	8			
Science	40	40	25			
C Life	11	11	8			
C Physical	11	11	6			
C Earth	11	11	6			
C Nature of Science	7	7	5			
P Models	14	14	10			
P Constancy	13	13	9			
P Form & Function	13	13	6			

Clusters	NP	NA	NC	Below Avg	Avg	Above Avg
Science (cont.)						
P Thinking Skills	20	20	13			
Social Science	40	40	29			
C History	10	10	7			
C Geography	10	10	8			
C Political Science	10	10	8			
C Economics	10	10	6			
P App. of Knowledge/Comp.	16	16	11			
P Org., Summ. & Interp. of Info.	13	13	9			
P Determination of Cause/Effe ct	11	11	9			
P Thinking Skills	19	19	13			
Listening	40	40	31			
C Vocabulary	10	10	8			
C Comprehension	30	30	23			
P Initial Understanding	11	11	5			
P Interpretation	12	12	12			
P Analysis	7	7	6			
C Literary	10	10	7			
C Informational	10	10	7			
C Functional	10	10	9			
P Thinking Skills	19	19	18			
Thinking Skills	174	174	131			

STANFORD LEVEL/FORM: PRIMAR Y 3/A
2002 NORMS: Spring National

OLSAT LEVEL/FORM: D/5
2002 NORMS: Spring National

C = Content Cluster P = Process Cluster

Scores based on normative data copyright © 2003 by Harcourt Assessment, Inc. All rights reserved

COPY 01
PROCESS NO. 10303657-2461655-1409-01465-1

Figura 11-2. Informe muestra de puntuaciones del SAT10.
Fuente: Stanford Achievement Test, Décima Edición, informe muestra de un estudiante.
Copyright © 2003 por Harcourt Assessment, Inc.
Reproducido con autorización. Todos los derechos reservados.

Los programas de investigación del SAT10 incluyeron cerca de 170 000 estudiantes en la fase de prueba de reactivos, 250 000 estudiantes de 650 distritos escolares en el programa de estandarización de primavera, 110 000 en el de otoño y 85 000 estudiantes en distintos programas de igualación. Este extraordinario esfuerzo se dedicó a asegurar la representatividad de estas muestras en términos de distribución racial/étnica, estatus

socioeconómico, región geográfica y tipo de escuela (pública/privada). El manual técnico describe con cuidado la naturaleza de estos programas de investigación y las características de los participantes.

El cuadro 11-5 presenta los datos de confiabilidad de las principales puntuaciones de una batería del SAT10. Aparecen consistencia interna (KR-20 para la Forma S) y formas alternas (Formas S y T). Como suele suceder con estos tipos de pruebas, la confiabilidad de formas alternas tiene un promedio entre .05 y .08 menor que la confiabilidad de consistencia interna. Los datos ilustran la relación común entre el número de reactivos de la prueba y el nivel de confiabilidad. Cuando el número de reactivos es mayor de 60, la confiabilidad de la consistencia interna tiende a estar alrededor de .95; cuando está entre 30 y 40, se ubica en el rango de .85 a .90; y cuando es menor de 10, como en el caso de las puntuaciones de grupos que aparecen en algunos informes, las confiabilidades, por lo general, son muy bajas. Como señalamos en el capítulo 4, se debe tener cuidado con las pruebas cortas.

Cuadro 11-5. Datos de confiabilidad de las pruebas del SAT10 en la batería P3

Prueba	No. de reactivos	KR-20	Forma alterna
Lectura total	114	.96	.90
Habilidades para estudiar palabras	30	.85	.82
Vocabulario de lectura	30	.87	.82
Comprensión de lectura	54	.93	.83
Matemáticas total	76	.94	.90
Matemáticas: Resolución de problemas	46	.91	.85
Matemáticas: Procedimientos	30	.90	.83
Lenguaje	48	.91	.86
Deletreo	38	.88	.80
Ciencia	40	.86	.80
Ciencia social	40	.89	.79
Habilidades de escuchar	40	.84	.75

Fuente: Adaptado de *Stanford Achievement Test*, Décima Edición, *Spring Technical Data Report*, cuadros C-4 y E-3.

Usos técnicos y características especiales

Las baterías de aprovechamiento se usan para una gran variedad de propósitos, lo cual a veces se vuelve problemático, porque no siempre son compatibles dichos propósitos. La intención original de estas pruebas era monitorear el progreso de los estudiantes individuales en las áreas importantes del plan de estudios escolar, y el maestro era el principal receptor de esta información. Éste es aún un propósito importante y, quizá, la aplicación más común de estas pruebas; sin embargo, otros usos también se han extendido. Las escuelas y los distritos escolares, por ejemplo, ahora emplean resúmenes

de las puntuaciones de grupos de estudiantes para evaluar el plan de estudios. En algunos casos, estos resúmenes también se usan para evaluar la eficacia del personal de la escuela: maestros, directores y superintendentes. Este uso, desde luego, es muy diferente del propósito original.

Ahora, las puntuaciones de las baterías de aprovechamiento se informan de manera cotidiana a los padres. Además, los resúmenes grupales se informan a los comités escolares y a las comunidades locales como medidas de la eficacia escolar. Por último, las baterías de aprovechamiento (o partes de ellas) se emplean en muchos proyectos de investigación como variables dependientes. Por ejemplo, los resultados de lectura y matemáticas del Stanford en los grados de primaria pueden usarse como medidas del efecto de la participación en un programa de estancia infantil. Estos usos otorgan un peso extraordinario a los requerimientos de desarrollo y los materiales interpretativos de estas pruebas.

Los materiales de las editoriales de las baterías de aprovechamiento de manera comprensible hacen hincapié en sus características distintivas. Sin embargo, incluso una observación casual revela que las semejanzas entre estas baterías superan en cantidad a las diferencias. Identifiquemos algunas características en común.

Primero, aunque una de estas baterías puede denominarse como "una" prueba, por ejemplo, la prueba de Stanford o la de Iowa, en realidad constituyen un sistema de muchas pruebas interrelacionadas. Existe una docena o más de pruebas separadas de cada nivel, de niveles múltiples y de formas múltiples. Cada vez más hay versiones largas y cortas, versiones de opción múltiple y de respuesta abierta, y ediciones en distintos idiomas y otras variaciones por el estilo. Además, puede haber bancos de reactivos, versiones personalizadas y ediciones seguras de las series. Las versiones modernas de estas baterías, en realidad, constan de más de 100 pruebas separadas identificables; sin embargo, todas estas pruebas están relacionadas entre sí, no sólo por el nombre, sino en términos de procedimientos de elaboración, estructuras normativas y sistemas interpretativos.

Segundo, aparte del número de pruebas separadas identificables, el conjunto de materiales complementarios e informes de puntuaciones de estas baterías es asombroso. Puede haber cuadernillos interpretativos separados para estudiantes, padres, maestros y administradores escolares. Hay numerosas opciones para los informes generados por computadora de las puntuaciones. Existen en el mercado pruebas prácticas, pruebas de localización, listas detalladas de objetivos y una gran cantidad de otros materiales auxiliares.

Tercero, los procedimientos de estandarización y otros programas de investigación de las baterías de aprovechamiento son, en general, ejemplares, pues emplean las metodologías más recientes y sofisticadas. Las muestras suelen ser buenas de acuerdo con los requerimientos de la estabilidad estadística. Se toman medidas extraordinarias para asegurar la representatividad por género, raza, grupo étnico, región geográfica, estatus socioeconómico y otras características demográficas. Los manuales técnicos de estas baterías son exhaustivos.

Cuarto, las baterías de aprovechamiento han sido criticadas por mucho tiempo debido a que se basan de manera exclusiva en la metodología de opción múltiple, pero esto ya no es así. Todas las baterías importantes de aprovechamiento emplean ahora métodos de evaluación además de los reactivos de opción múltiple. Por ejemplo, ahora todas tienen alguna forma de ejercicio de escritura libre para la evaluación de la calidad de la escritura. Algunas baterías importantes ofrecen conjuntos extensos de medidas de respuesta abierta y de ejecución. Los costos para calificar estas formas de evaluación son muy altos, lo cual limita su uso real.

Quinto, todas las baterías de aprovechamiento dependen en gran medida de las *mismas fuentes de información de su contenido*, entre las que se encuentran a) afirmaciones de las metas curriculares de organizaciones profesionales como el *National Council of Teachers of Mathematics* (NCTM), *National Council of Teachers of English* (NCTE) y organizaciones similares, b) series de libros de texto importantes, c) resúmenes preparados por organizaciones como *National Assessment of Educational Progress* (véase más adelante en este capítulo), d) estándares de contenido de los departamentos estatales de educación, en donde los estados con mayor población tienen particular influencia en este aspecto; y e) los nuevos *Common Core State Standards* [Estándares Estatales Básicos en Común], aunque están limitados por el momento a las matemáticas y las letras inglesas. Al establecer el contenido de estas pruebas, para bien o para mal, no encontrarás "disidentes" tomando nuevas direcciones.

Baterías de aprovechamiento de nivel universitario

Ahora existen varias baterías diseñadas para su uso en la universidad, las cuales se concentran en los resultados educativos generales de los programas de licenciatura en áreas como desarrollo de la habilidad de escritura, capacidad para usar la computadora y encontrar información, y elementos, al menos generales, de las humanidades, ciencias naturales y ciencias sociales.

Esto contrasta con el campo principal de estudio (p. ej., biología, psicología, enfermería o contaduría) que podría ser abordado por alguna de las pruebas que revisaremos más adelante en el apartado de pruebas de aprovechamiento de área única y los exámenes de licencia o certificación.

El cuadro 11-6 presenta tres ejemplos de baterías de aprovechamiento de nivel universitario que, aunque se discuten mucho, no se emplean tanto en las universidades. Sin embargo, su uso ha aumentado un poco en los años recientes al mismo tiempo que el movimiento de responsabilidad se ha extendido en las universidades. Se ha aprobado el uso de estas tres baterías por una iniciativa reciente llamada *Voluntary System of Accountability* [Sistema Voluntario de Responsabilidad] (www.voluntarysystem.org/). Es difícil estimar el futuro de este intento.

Cuadro 11-6. Ejemplos de baterías de aprovechamiento de nivel universitario

Título	Editorial
Collegiate Assessment of Academic Proficiency (CAAP)	ACT, Inc.
ETS Proficiency Profile (EPP) [a]	ETS
Collegiate Learning Assessment (CLA)	*Council for Aid to Education*

[a] Antes conocido como *Measure of Academic Proficiency and Progress* (MAPP), el cual a su vez reemplazó al *Academic Profile* (AP).

Pruebas de aprovechamiento de área única

Existe una gran variedad de pruebas de aprovechamiento que cubren un dominio de contenido único. Por lo general, están diseñadas para usarse en bachillerato o la universidad, a menudo al final de un curso o un programa entero de estudio, por ejemplo, un programa de formación técnica o vocacional de una universidad importante. Estas pruebas también están disponibles como pruebas de competencia ocupacional. En el esquema de clasificación que hemos adoptado en este capítulo, excluimos de esta categoría las *partes* de baterías de aprovechamiento (p. ej., la prueba de matemáticas de la *Prueba de Aprovechamiento de Stanford*) que pueden emplearse como pruebas por sí mismas. También excluimos las pruebas relacionadas con ocupaciones por medio de las cuales se obtiene una licencia o certificación. Es evidente que estos tipos de pruebas pueden entrar en la categoría de pruebas de aprovechamiento de área única, pero hablaremos de ellas en otra sección de este capítulo.

¡Inténtalo!

Para ilustrar el vasto número de pruebas de la categoría de pruebas de aprovechamiento de área única, visita la página de internet de ETS Test Collection (http://www.ets.org/test_link/find_tests/). Escribe el nombre de un campo de conocimiento, como biología, psicología, español o matemáticas [biology, psychology, Spanish, mathematics] y revisa la lista de resultados. Observa que, además de las pruebas relacionadas con el área, probablemente encontrarás medidas de actitud hacia dicha área.

Ejemplos

Consideremos algunos ejemplos para ilustrar esta enorme categoría de pruebas. El primer ejemplo es el *Major Field Test in Psychology* [Prueba de Áreas Importantes: Psicología] (MFT-P, Forma 4GMF). Ésta es una de una serie de pruebas diseñadas para medir el aprendizaje del alumno en un campo de estudio universitario. También hay pruebas disponibles en áreas como química, economía e historia. De acuerdo con los creadores de las pruebas, éstas "evalúan el dominio de conceptos, principios y conocimiento esperado de los alumnos que han terminado, o están a punto de hacerlo, un programa de estudios. Cada prueba es elaborada por un panel nacional de expertos en el tema y se basa en el contenido central de los planes de estudio que se identifican en una investigación de programas nacionales" (*Educational Testing Service,* 2012, p. 1). En su origen, las pruebas se diseñaron para ser versiones más cortas y menos difíciles de las pruebas correspondientes de los temas de Graduate Record Examinations (GRE). Se pueden contestar en línea o en formato de lápiz y papel; la versión en línea no es una prueba adaptada para computadora, sino que presenta simplemente la prueba de manera electrónica.

El MFT-P consta de 140 reactivos de opción múltiple que se aplican en dos sesiones.

Por cada examinado, la prueba produce una puntuación total y cuatro subpuntuaciones, que se enumeran en el cuadro 11-7, el cual también muestra los datos de confiabilidad. Observamos de nuevo la relación entre el número de reactivos y la confiabilidad: muchos reactivos, buena confiabilidad, y pocos reactivos, no muy buena confiabilidad. Además, aparecen "indicadores de evaluación" en seis áreas por grupos de examinados; estos indicadores se basan en grupos relativamente pequeños de reactivos que, con claridad, no producirían puntuaciones confiables de los individuos.

Cuadro 11-7. Esquema de las puntuaciones del Major Field Test – Psychology II

Puntuaciones individuales	**Reactivos**	**Confiabilidad**[a]
1. Aprendizaje y cognición	27	.73
2. Percepción y sensación, fisiología, Psicología comparada, etología	22	.68
3. Psicología clínica, Psicología anormal, personalidad	25	.69
4. Psicología social y del desarrollo	31	.78
Total	140[b]	.93

[a] La confiabilidad KR-20 se basa en los alumnos del último año (*Educational Testing Service*, 2012). El número de reactivos para la confiabilidad difiere ligeramente del número de reactivos de la prueba presentados.
[b] El número de reactivos de las subpuntuaciones no suman 140, ya que algunos reactivos no entran en ninguna de las cuatro subpuntuaciones, pero sí entran en la puntuación total.

El manual proporciona normas de percentiles basadas en individuos y promedios institucionales. (En la página XX se puede encontrar la distinción entre estos dos tipos de percentiles). Las normas derivan de las escuelas que han usado la prueba en años recientes, por lo que es un ejemplo clásico de las normas del usuario. El manual hace hincapié en que se trata de normas del usuario más que de normas representativas nacionales. Las normas de 2012 se basan en 4603 alumnos del último grado y 167 instituciones.

Las pruebas elaboradas por el *National Occupational Competency Testing Institute* (**NOCTI**) ofrecen una enorme cantidad de ejemplos de pruebas de aprovechamiento de área única. NOCTI tiene más de 150 pruebas de competencia ocupacional para nivel de ingreso y trabajadores experimentados en campos como fabricación de dados, reparación de aparatos, cosmetología y programación de computadoras. Una prueba típica de NOCTI tiene cerca de 180 reactivos de opción múltiple y un tiempo de aplicación de 3 hrs. Algunas también tienen un componente de ejecución; cada prueba se basa en un esquema detallado de contenido de las habilidades y conocimientos que se consideran importantes para la ocupación. Cada prueba tiene "normas del usuario" basadas en todos los examinados que contestaron la prueba en los años recientes. En el caso de algunas pruebas de NOCTI, el número se reduce a unos cuantos casos y, en otras, el número asciende a varios miles.

¡Inténtalo!

Revisa la página de internet de NOCTI (nocti.org) para ver el conjunto de pruebas. Observa las especificaciones del contenido (anteproyectos) que se enumeran de cada prueba. Estas especificaciones se relacionan con la validez de contenido.

Un ejemplo final de las pruebas de aprovechamiento de área única es la prueba STAR Math. Aquí describimos la versión "clásica". Se trata de una **prueba adaptada paracomputadora** que se usa en los grados 1 al 12. El fondo de reactivos consta de cerca de 2400 reactivos de opción múltiple que van desde los problemas de suma simple hasta los de álgebra y geometría de bachillerato. Un estudiante individual responde sólo 24 reactivos que elige el programa de cómputo; el primer reactivo se elige de acuerdo con el grado del alumno. Después, una respuesta correcta lleva a un reactivo más difícil, mientras que una incorrecta lleva a un reactivo más fácil. Al final se espera que pueda determinarse con exactitud la ubicación del alumno. Los datos normativos nacionales del STAR Math producen equivalentes de grado, equivalentes de curva normal y rangos percentiles. Un informe narrativo generado por computadora proporciona la recomendación respecto de la ubicación instruccional con referencia a un criterio. La puntuación única de la prueba tiene confiabilidades de formas alternas dentro del grado que van de .72 a .80 con una mediana de .74, y confiabilidades de división por mitades que van de .78 a .88 con una mediana de .85 (Renaissance Learning, 2003). Un creciente número de pruebas adaptadas para computadora como STAR Math está disponible para usarse en escuelas.

¡Inténtalo!

Para ver ejemplos de reactivos e informes del STAR Math, visita esta página de internet: http://www.renlearn.com/sm/

Usos típicos y características especiales

Las pruebas de aprovechamiento de área única tienen dos usos típicos. Primero, se usan para determinar el desempeño del individuo en un área sumamente focalizada, como un cuerpo de conocimientos o habilidades. Segundo, pueden usarse para evaluar lo adecuado de un programa de enseñanza; en este caso se supone no sólo que la prueba tiene validez de contenido, sino también que los individuos hacen un esfuerzo razonable para adquirir el conocimiento o habilidad que se evalúa.

¿Qué generalizaciones se pueden hacer acerca de las pruebas de aprovechamiento de área única? Hay un conjunto tan vasto de estas pruebas que hacer generalizaciones es un poco arriesgado; sin embargo, al menos si nos limitamos al ámbito de los ejemplos típicos que ofrecen las editoriales importantes, podemos discernir algunas características en

común.

Casi todas las pruebas de esta categoría son de aplicación grupal, de lápiz y papel y de opción múltiple. Sin embargo, en los últimos 10 años, cada vez hay más pruebas disponibles en versiones computarizadas.

Algunas pruebas relacionadas con ocupaciones también tienen un componente de ejecución. La extensión de la prueba tiende a estar en un rango de 50 a 200 reactivos, pero muchas tienen más de 100. Las pruebas adaptadas para computadora tienen grandes fondos de reactivos, pero el examinado, en realidad, contesta una cantidad relativamente pequeña de ellos. La mayoría de las pruebas de esta categoría hace hincapié en el uso de puntuaciones únicas totales aunque sigue produciendo algunas subpuntuaciones. Las puntuaciones totales tienden a ser muy confiables, por lo general de .90 o mayores, lo cual no es sorprendente. El nivel de conocimiento de la gente de un campo de información bien definido, conceptos y habilidades, es un fenómeno bastante estable cuando se mide con una muestra grande de reactivos. La confiabilidad de las subpuntuaciones es casi siempre una función del número de reactivos de la subprueba. Como regla general, debemos ser cautelosos con la confiabilidad de cualquier puntuación basada en menos de 20 reactivos.

Para indicar la validez, las pruebas de esta categoría dependen casi exclusivamente de la validez de contenido. El contenido de la prueba se compara con un curso individual o un programa entero, o con una descripción sistemática del puesto. Hay intentos ocasionales para reunir otros tipos de evidencia de la validez de estas pruebas, pero predomina la validez de contenido. Debido a que cambian los contenidos de los programas de estudio o las especificaciones del puesto, estas pruebas tienden a actualizarse de manera regular; es raro encontrar una de ellas en uso práctico que tenga más de 10 años de antigüedad.

Pruebas de licencia y certificación

Consideramos aquí las pruebas que tienen como finalidad conceder una licencia o certificación. Existen numerosas pruebas en esta categoría. A veces, hemos sido descritos como una "sociedad credencializada", y las pruebas tienen un papel destacado, aunque sin duda no exclusivo, en este proceso. Los requerimientos educativos también tienen un papel importante; a veces cumplir con ellos es suficiente. Sin embargo, a menudo ocurre que una persona debe pasar algún tipo de examen antes de obtener la licencia o certificación. La preparación, aplicación e interpretación de las puntuaciones de estos exámenes, por lo general, siguen los principios bosquejados en secciones anteriores de este libro. Agrupamos los exámenes de licencia y certificación porque tienden a ser muy parecidos en sus propósitos, estructura, elaboración y uso. Sin embargo, existe una distinción técnica entre unos y otros. La **licencia** implica una concesión legal por parte de una agencia gubernamental para ejercer una profesión; la mayoría de las licencias está bajo el control del estado, aunque algunas son concedidas por el gobierno federal. El principal interés de la licencia es determinar que el solicitante posea el nivel mínimo de conocimientos y/o habilidades para proteger al público usuario. La **certificación** implica la declaración de que se ha alcanzado un nivel de eficiencia, pero esto no necesariamente implica un derecho a algo. La certificación puede buscar un nivel mínimo o superior de conocimientos o habilidades dependiendo del propósito particular.

Ejemplos

El examen que se presenta para tener licencia como terapeuta médico constituye un ejemplo. El *National Physical Therapy Examination* es controlado por la *Federation of State Boards of Physical Therapy* (véase www.fsbpt.org . La aplicación real es manejada por medio de la red nacional de centros de evaluación psicométrica (véase www.prometric.com).

El examen consta de 250 reactivos de opción múltiple que se tienen que responder en un periodo de 4 horas. Aunque la prueba se aplica por computadora, no está adaptada para la versión computarizada; es decir, la computadora presenta los 250 reactivos en la misma sucesión al examinado en vez de elegir los reactivos con base en las respuestas a los reactivos previos. El contenido de la prueba sigue un **anteproyecto de la prueba** o esquema de contenido que está ya disponible para los candidatos (véase el sitio fsbpt.org para consultar el esquema de contenido). Las puntuaciones naturales de la prueba se convierten en puntuaciones estándar. La Federation of State Boards of Physical Therapy ha establecido una sola puntuación estándar como aprobatoria que se adoptó en todos los estados. Este examen de terapia médica es típico de las pruebas de licencia y certificación.

Un segundo ejemplo es el examen para ser piloto privado certificado, que se aplica con autorización del *Federal Aviation Administration* (FAA), el cual es responsable de

conceder licencias a todos los pilotos de EUA. Podemos notar que, mientras que el título dice "certificado", es técnicamente una licencia según la definición que dimos antes. El examen incluye preguntas escritas (de opción múltiple) y un examen de vuelo. El examinado debe tener un tiempo de vuelo a solas, cumplir con el requerimiento de la edad, contar con un certificado médico y tener competencia en la lengua inglesa. Una persona debe pasar todas las partes del examen para recibir la certificación como piloto. En realidad, hay una gran variedad de distintas licencias para distintos tipos de naves (helicópteros, globos aerostáticos, etc.) y circunstancias (comercial, no comercial, etc.). La licencia de piloto es una de las pocas que se controla a nivel federal y no a nivel estatal. Podríamos señalar que algunos de los exámenes de NOCTI mencionados antes también son usados por organizaciones profesionales con propósitos de certificación.

Usos típicos y características especiales

Poco se necesita decir para describir los usos típicos de las pruebas de licencia y certificación. Se usan para documentar, de una manera muy oficial, la competencia del examinado en habilidades y conocimientos pertinentes. Sin embargo, existe un uso secundario cuando los examinados son producto de un programa educativo claramente identificable. En esos casos, el administrador del programa usa los resultados de la prueba como medida de la eficiencia del programa. Por lo general, varios tipos de asociaciones profesionales supervisan el uso de las pruebas de esta categoría, pero a menudo contratan a organizaciones profesionales de evaluación para preparar las pruebas.

Las características distintivas de las pruebas de licencia y certificación surgen de las consecuencias cruciales que tienen para los examinados. Primero, con el fin de producir puntuaciones confiables, estas pruebas tienden a ser bastante largas. Son comunes las pruebas que tienen varios cientos de reactivos y que requieren varias horas para su aplicación. Como resultado, las puntuaciones totales de estas pruebas tienden a ser muy confiables, por lo común, alrededor de .95. Segundo, las pruebas suelen tener esquemas de contenido explícitos y muy claros. Los examinados, por lo general, tienen acceso a estos esquemas antes de presentar la prueba. Tercero, una extraordinaria seguridad rodea estas pruebas, pues pueden ser aplicadas sólo una o pocas veces al año en lugares cuidadosamente controlados. Además, para proteger la seguridad del contenido de la prueba, se cambian con frecuencia los reactivos. En el caso de algunos tipos de evaluación, puede ser una virtud usar los mismos reactivos en varias ocasiones, pero no sería prudente usar exactamente los mismos reactivos de manera repetida en un examen de certificación. Cuarto, los creadores de pruebas y quienes las contestan, de manera comprensible, se preocupan por los puntos de corte y cómo se obtienen. Casi siempre, un buen elemento de juicio está implicado en fijar las puntuaciones de corte, aunque los juicios se basan en el desempeño real de los examinados.

Cómo establecer puntuaciones de corte

Antes señalamos el papel crucial de las puntuaciones de corte para los exámenes de licencia y certificación. El tema también es importante para las pruebas de competencia que existen en los programas estatales de evaluación; en estos casos, un alumno debe obtener cierta puntuación para cumplir un requisito de graduación o promoción. Mencionamos este uso en nuestra descripción de un programa típico de evaluación escolar, así como al hablar de los programas estatales de evaluación. Además, muchas pruebas ofrecen alguna interpretación con referencia a un criterio que traduce el desempeño a categorías como básico, competente y avanzado, o a un esquema similar. En todos estos casos, está la pregunta de cómo se determina la puntuación de corte: el punto que separa entre aprobar y reprobar, competente e incompetente, o un nivel de eficiencia y otro nivel. Se ha dedicado mucha atención a esta pregunta en la literatura psicométrica. Establecimiento de estándares [en inglés, standard setting] es la etiqueta oficial que se aplica a este tema, aunque también se aplica el término puntuaciones de corte o puntos de corte. En Cizek (2001b) y Cizek y Bunch (2007) se puede encontrar un tratamiento detallado de este tema.

Los métodos para tratar el problema de establecer puntuaciones de corte caen en dos categorías principales: con referencia a una norma y con referencia a un criterio. En el primer método se establece la puntuación de corte por referencia a algún punto en un conjunto de normas; por ejemplo, "competente" puede definirse como ubicarse en el 25% superior, es decir, en el percentil 75 o mayor de acuerdo con las normas de la prueba, u obtener un equivalente de grado de al menos 8.0. Este método es común en la evaluación para contratar personal de trabajo; por ejemplo, los 10 mejores candidatos se citan para entrevistarlos. Este método tiene una atractiva sencillez; además, el hecho de que se basa en el desempeño real en la prueba tiene mérito. En algunas situaciones de mucha competencia, éste puede ser el único método disponible.

Por otro lado, un enfoque estrictamente con referencia a una norma tiene claros inconvenientes en algunos contextos. Consideremos el caso de una prueba de competencia mínima; puede ser que todos hayan alcanzado el nivel requerido de competencia. No hay razón para que alguno repruebe, ni siquiera el 10% inferior; o puede ser que nadie haya alcanzado el nivel requerido de competencia. Estos son juicios con referencia a un criterio. Nos gustaría tener un conjunto de procedimientos para examinar el contenido de la prueba y decidir qué tan bueno debe ser el desempeño del examinado para que sea declarado competente, aprobado o aceptable. Se han propuesto numerosos métodos para ayudar a formular tales juicios; Cizek y Bunch (2007) ofrecen útiles resúmenes de los métodos. En Chinn y Herz (2002) se puede encontrar un ejemplo de aplicación práctica en la certificación de terapeutas de pareja y familiares. En los años recientes, hemos sido testigos de la proliferación de métodos para establecer puntuaciones de corte, sobre todo en relación con los programas estatales y nacionales de evaluación que se describen en la siguiente sección. Gran parte de la investigación se ha

destinado a las consecuencias de aplicar estos métodos; quizá el más popular de ellos es el de Angoff o sus modificaciones. En esencia, el procedimiento implica imaginar una "persona mínimamente competente" y hacer un juicio por cada reactivo acerca de la probabilidad de que esta persona obtenga el reactivo correcto; después, se suman las probabilidades de todos los reactivos. El resultado será la puntuación de corte, es decir, la puntuación obtenida por la persona mínimamente competente. Este procedimiento también se puede usar imaginando otras personas, por ejemplo, una persona "sumamente competente". ¿Quién hace los juicios? En muchas aplicaciones, expertos en la materia.

En la práctica, los dos métodos (con referencia a una norma y con referencia a un criterio) suelen trabajar en conjunto. En el método empírico con referencia a una norma, una persona o un grupo debe hacer un juicio de que el percentil 75 es apropiado en lugar de uno de 60 o 90. Así, aunque la puntuación de corte se define de manera empírica, un juicio determina el punto exacto elegido. En el enfoque crítico con referencia a un criterio, antes de que se emita un juicio final, casi siempre ocurre que los jueces observan los resultados empíricos de la evaluación, a la luz de los cuales a menudo se modifican los juicios originales.

¡Inténtalo!

He aquí un conjunto sencillo de reactivos de aritmética.
$13 \times 29 =$ _____ $462/18 =$ _____
$.07 \times 3 =$ _____ $.28/.009 =$ _____
¿Cuántos reactivos debe tener correctos un estudiante de cuarto grado para que se considere "competente" en matemáticas?
¿Cuántos reactivos debe tener correctos un estudiante universitario de primer año para que se considere "competente" en matemáticas?

Pruebas de aprovechamiento estatales, nacionales e internacionales

Incluimos en esta categoría todas las pruebas de aprovechamiento que están bajo control de un grupo gubernamental o cuasi-gubernamental, o son financiadas por éste. Desde algunos puntos de vista, la clasificación es cómoda, aunque la organización no es por completo satisfactoria como será evidente conforme avance nuestra descripción.

Programas estatales de evaluación

Aunque algunos estados contaron con programas de evaluación por muchos años, la ley *No Child Left Behind* (**NCLB**] ordenó a todos los estados llevar a cabo estos programas si querían recibir dinero del erario federal. A pesar de que las pruebas difieren en cada estado, tienen algunas características en común; por ejemplo, todas abordan lectura/lenguaje, artes, matemáticas y ciencia en los grados 3-8 y al menos un grado en el rango 9-12. Muchos estados también cubren otras áreas curriculares.

La ley NCLB dicta el uso de puntuaciones de corte con referencia a un criterio para definir "eficiencia" y el uso de ciertas categorías para los informes. El contenido exacto de las pruebas difiere en cada estado, pues cada uno define el contenido y la manera exacta en que la prueba examinará dicho contenido. Cada estado también define sus propios niveles de eficiencia. Estas pruebas tienen una influencia excepcional en la manera de llevar a cabo la educación en los niveles de primaria y secundaria en EUA.

En 2011 y 2012 aparecieron dos desarrollos importantes que pueden influir en los programas estatales de evaluación. Primero, la gran mayoría de estados aprobó un documento llamado *Common Core State Standards* como definición de sus programas de estudios en lugar de que cada uno tenga su propia definición. ¿Acaso están muy atrasadas las pruebas de núcleo común? Ya veremos. Segundo, el gobierno federal empezó a facilitar la exención de requerimientos para la ley NCLB, la cual fue el origen de los programas estatales actuales de evaluación. Es difícil predecir cómo afectarán estas exenciones los programas de evaluación.

Un programa nacional de evaluación: NAEP

Muchas pruebas, como el SAT, ACT, OLSAT, WAIS y las series Stanford, pueden llamarse de manera legítima pruebas "nacionales", pero existe una iniciativa que merece esta denominación de manera especial. Se trata del *National Assessment of Educational Progress* [Evaluación Nacional del Progreso Educativo] (**NAEP**), que no es sólo una prueba, sino un proyecto de evaluación financiado con fondos federales. NAEP se formuló en la década de 1960 con el propósito específico de proporcionar un conjunto de

puntos de referencia en relación con los logros educativos de los estadounidenses. En sus primeros años, incluía a estudiantes de escuelas como a muestras de adultos que no asistían a la escuela. En años más recientes, ha estado confinado por completo a muestras de las escuelas, sobre todo de los grados 4 de primaria, 2 de secundaria y 3 de bachillerato. Cada fase de evaluación se concentra en sólo una o dos áreas curriculares, como lectura, matemáticas, escritura o geografía. Por ejemplo, NAEP abarcó la historia de EUA en 2010 y lo hará de nuevo en 2014. En la actualidad, NAEP cubre lectura y matemáticas cada año non. En algunas ocasiones, las pruebas cubren los tres grados mencionados, y en otras, no. Se hacen esfuerzos especiales para asegurar la representatividad de las muestras de estudiantes de todo el país y, en algunos casos, hacer informes de nivel estatal, en los que se incluyen resultados por género y grupo étnico. NAEP no informa puntuaciones de estudiantes individuales. Incluye reactivos de opción múltiple (respuesta elegida) y de respuesta abierta (respuesta construida). Algunos de estos últimos reactivos se han convertido en modelos para los esfuerzos de evaluación en donde se les denomina reactivos "tipo NAEP". Los informes de NAEP a veces se refieren a sí mismos, quizá de manera un poco pretenciosa, como el "programa de informes de la nación", término que aparece en su página de internet dentro del sitio *National Center for Education Statistics* [Centro Nacional para la Estadística en Educación] (nces.ed.gov/nationsreportcard). Otro de estos programas, el *National Assessment of Adult Literacy* [Evaluación Nacional de Alfabetismo Adulto] (NAAL), ha complementado los resultados de NAEP (que se limitan a estudiantes en las escuelas) con medidas de la capacidad para leer y escribir entre los adultos.

¡Inténtalo!

¿Quieres saber cómo te fue en las pruebas en tu bachillerato, secundaria o primaria de tu estado? La ley NCLB exigió que los estados hicieran públicos los resultados de los programas de evaluación, por lo general a la vista en los propios edificios escolares. Aunque cada programa estatal tiene un nombre diferente, es fácil encontrar el programa de tu estado escribiendo "programa de evaluación de [nombre del estado]" en cualquier buscador de internet. Hazlo y encontrarás una gran cantidad de información acerca de los resultados de tu estado.

¡Inténtalo!

Entra a la página de internet de NAEP (nces.ed.gov/nationsreportcard). Luego consulta los reactivos muestra de una de las áreas curriculares que se evalúan. ¿Qué clase de reactivos encontraste? También, observa los métodos que se emplearon para informar los resultados de muestras nacionales y subgrupos.

Programas internacionales de evaluación: TIMSS, PIRLS y PISA «301a

Desde sus más bien oscuros inicios en 1964, el proyecto que se conoce en la actualidad como *Trends in International Mathematics and Science Study* [Estudio de las Tendencias Internacionales en Matemáticas y Ciencia] (TIMSS) empezó evaluando a estudiantes de entre 30 y 40 países, la mayoría perteneciente a las economías más desarrolladas del mundo. La evaluación se concentró en ciencia y matemáticas con estudiantes de los grados 4 de primaria y 2 de secundaria (o sus equivalentes de cada país). En EUA se anunciaba cada vez más que serían los "primeros en el mundo" o que "cumplirían con los estándares de clase mundial"; los resultados de TIMSS, al principio muy ignorados en EUA, han atraído una considerable atención del público. Al igual que NAEP, TIMSS no informa resultados de estudiantes individuales, sino que produce resúmenes grupales, en este caso, puntuaciones promedio de países enteros. En años más recientes, el *Program for International Student Assessment* [Programa para la Evaluación Internacional de Estudiantes] (PISA) ha tomado como meta a estudiantes de 15 años de numerosos países en las áreas de lectura, matemáticas y ciencia. PISA afirma que hace hincapié en el "alfabetismo funcional" más que en el aprendizaje escolar. Otra prueba en la siempre creciente lista de evaluaciones internacionales es el *Progress in International Reading Literacy Study* [Progreso en el Estudio Internacional de la Lectura] (PIRLS), que cubre el área de lectura del grado 4 de primaria. Todos estos programas tienen un calendario periódico para reunir datos: TIMSS, cada cuatro años (2007, 2011, ...); PISA, cada tres años (2009, 2012, ...); PIRLS, cada cinco años (2006, 2011, ...).

Cada uno de los programas nacionales e internacionales que identificamos aquí tiene una excelente página de internet que ofrece informes detallados de resultados, reactivos ilustrativos, descripciones de los planes de muestreo y así sucesivamente. Además, el *National Center for Education Statistics* informa muchos de los resultados clave de estos proyectos en una publicación anual llamada *The Condition of Education*, que también tiene un sitio de internet accesible. Todas las páginas pertinentes se enumeran en el Resumen de puntos clave 11-1.

Características especiales

Los programas estatales de evaluación, aunque controlados de manera separada por cada estado, se han vuelto muy similares desde la aprobación de NCLB. Es difícil predecir qué sucederá con la gradual erosión de NCLB. Los programas pueden seguir siendo similares, diferir o evolucionar en un solo programa nacional basado en los *Common Core State Standards*. Los esfuerzos nacionales e internacionales de evaluación comparten características distintivas. Primero, el énfasis de estos programas se pone en el desempeño grupal, pues no intentan informar puntuaciones de los individuos. Segundo, cada ola de evaluación a menudo aborda o hace hincapié en una sola área, por ejemplo, lectura o matemáticas. Incluso cuando se cubren dos áreas en un solo año, se trata de proyectos en gran parte independientes. En contraste, una batería de aprovechamiento

produce puntuaciones de una docena de distintas áreas por cada estudiante. Tercero, los programas nacionales e internacionales de evaluación se concentran en un pequeño número de grados o edades. Cuarto, casi todas estas pruebas incluyen reactivos de respuesta elegida (opción múltiple) y construida. Por último, estos programas intentan reunir una amplia información acerca de las prácticas educativas (p. ej., cantidad de tareas en casa, tiempo de enseñanza, etc.) que podrían ayudar a interpretar los resultados.

Resumen de puntos clave 11-1

Programas nacionales e internacionales importantes de evaluación
NAEP http://nces.ed.gov/nationsreportcard/
TIMSS http://nces.ed.gov/timss/ o
http://timss.bc.edu/
PISA http://www.oecd.org/pisa/ o
http://nces.ed.gov/surveys/pisa/
NAAL http://nces.ed.gov/naal/
PIRLS http://timssandpirls.bc.edu/

Pruebas de aprovechamiento de aplicación individual

Todas las pruebas de aprovechamiento que hemos presentado hasta ahora son de aplicación grupal. Sin embargo, existen algunas que se diseñaron explícitamente para ser aplicadas de manera individual, casi de la misma manera que las pruebas de capacidad que vimos en el capítulo 8.

Ejemplos

Veremos dos ejemplos de las pruebas de aprovechamiento de aplicación individual. Primero, está la *Prueba de Aprovechamiento Individual de Wechsler*, Tercera Edición (en inglés, Wechsler Individual Achievement Test, WIAT-III), que incluye 16 subpruebas en áreas como Lectura de palabras, Composición de ensayos, Deletreo y Fluidez matemática–Suma, y ocho puntuaciones compuestas en áreas como Lenguaje oral, Expresión escrita, Lectura básica, Matemáticas y Aprovechamiento total. La prueba proporciona normas para edades de 4 a 51 años. El principal objetivo del WIAT es investigar las discrepancias capacidad-aprovechamiento y las diferencias entre las puntuaciones de subpruebas.

La aplicación del WIAT es similar a la del WISC. El examinador se sienta frente al examinado, con el cuadernillo de estímulos sobre la mesa. El examinador establece el *rapport* y realiza la aplicación con un tono de conversación, pero con dominio de la presentación de las instrucciones estandarizadas. El examinador debe tener una formación amplia en el uso del WIAT para asegurar que la aplicación sea válida. Al igual que con WISC y WAIS, existen puntos de inicio, reglas de discontinuación y rúbricas de calificación que se aplican en el curso de la prueba. El tiempo típico de aplicación va de media ahora, en el caso de las edades más jóvenes, hasta más de hora y media, en el caso de las edades mayores.

Las normas del WIAT se basan en muestras elegidas con mucho cuidado de todo el rango de edades a las que está dirigida la prueba. Las puntuaciones naturales se pueden traducir en puntuaciones estándar (M = 100, DE = 15), equivalentes de edad o grado, rangos percentiles, estaninas, equivalentes de curva normal y puntuaciones estándar dcombinadas por edad y grado. Lo más importante es que las normas del WIAT tienen referencias cruzadas con varias escalas de inteligencia Wechsler (véase capítulo 8, pp. 209-213a»), las cuales constituyen la base para examinar las discrepancias capacidad-aprovechamiento. El manual ofrece cuadros con la magnitud de las diferencias entre varios pares de puntuaciones que se requiere para alcanzar ciertos niveles de significancia estadística. Un extenso informe del WIAT se orienta hacia los requerimientos y el lenguaje de un programa educativo individualizado (PEI) que exigen las leyes federales para personas con discapacidades. Por ejemplo, el informe contiene indicaciones sobre

cómo redactar las "metas anuales" y los "objetivos a corto plazo" relacionados con los reactivos y puntuaciones del WIAT. En Miller (2010) y Willse (2010) se pueden encontrar reseñas del WIAT-III.

El segundo ejemplo de las pruebas de esta categoría es la tercera edición del *Woodcock-Johnson* (WJ III; Woodcock, McGrew, & Mather, 2001). En sus ediciones previas, la prueba se conocía como *Batería Psicoeducativa de Woodcock-Johnson*. El WJ III es, en realidad, un sistema de dos pruebas: *Pruebas Woodcock-Johnson de Capacidad Cognitiva* (WJ III Cog) y *Pruebas Woodcock-Johnson de Aprovechamiento* (WJ III Ach). La figura 11-3 bosqueja las pruebas del WJ III Ach. La batería cognitiva, basada en la teoría de Cattell de la **inteligencia fluida** y **cristalizada** (los materiales del WJ III se refieren a la teoría de Cattell-Horn-Carroll o teoría CHC; véase la revisión de esta teoría en la página 182a»), tiene 10 pruebas en una batería estándar y 10 pruebas adicionales en una batería extensa. La batería de aprovechamiento tiene 12 pruebas en la batería estándar y permite hasta 10 pruebas adicionales en la batería extensa. En total, el sistema WJ III tiene 42 pruebas y numerosas puntuaciones compuestas derivadas de las combinaciones de estas pruebas, lo que conduce a un conjunto vertiginoso de información. La interpretación de la prueba, al igual que el WIAT, depende en gran parte de contrastar los niveles de desempeño entre esta multiplicidad de pruebas. El WJ III tiene normas para edades de 2 a 90 años obtenidas con muestras consideradas representativas de la población nacional de ese rango de edad. En Cizek (2003) y Sandoval (2003) se pueden encontrar revisiones del WJ III.

Área curricular	Batería estándar–Forma A o B	Batería extensa–Forma A o B
Lectura Habilidades básicas de lectura Fluidez de lectura Comprensión de lectura	Prueba 1: Identificación de letras-palabras Prueba 2: Fluidez de lectura Prueba 9: Comprensión de pasajes	Prueba 13: Ataque de palabras Prueba 17: Lectura de vocabulario
Lenguaje oral Expresión oral Comprensión auditiva	Prueba 3: Recuerdo de historias Prueba 4: Comprensión de instrucciones	Prueba 14: Vocabulario con dibujos Prueba 15: Comprensión oral
Matemáticas Habilidades de cálculo matemático* Fluidez matemática Razonamiento matemático	Prueba 5: Cálculo Prueba 6: Fluidez matemática Prueba 10: Problemas aplicados	Prueba 18: Conceptos cuantitativos
Lenguaje escrito Habilidades básicas de escritura Fluidez de escritura Expresión escrita**	Prueba 7: Deletreo Prueba 8: Fluidez de escritura Prueba 11: Muestras de escritura	Prueba 16: Redacción
Conocimiento académico		Prueba 19: Conocimiento académico
Complementario	Prueba 12: Recuerdo de historias–Demorado Escala de legibilidad de la escritura a mano Escala de evaluación de la escritura	Prueba 20: Deletreo de sonidos Prueba 21: Conciencia de los sonidos Prueba 22: Puntuación y uso de mayúsculas

= prueba en el cuadernillo de respuestast

= prueba cronometrada

= prueba grabada en audio

* Pruebas 5 y 6

** Pruebas 8 y 11

Figura 11-3. Bosquejo de las pruebas de *Woodcock-Johnson III Batería de Aprovechamiento*.

Fuente: Woodcock-Johnson® III Tests of Achievement Examiner's Manual (p. 3), con autorización de la editorial. Copyright © 2001 por Riverside Publishing Company. Todos los derechos reservados. Ninguna parte de las pruebas puede ser reproducida o transmitida en ninguna forma y por ningún medio, electrónico o mecánico, incluyendo fotocopias y grabación o por cualquier sistema de almacenamiento o recuperación de información sin autorización previa por escrito de The Riverside Publishing Company, a menos que tal copia esté permitida expresamente por la ley federal de derechos de autor. Dirige cualquier duda a Contracts and Permissions Department, The Riverside Publishing Company, 425 Spring Lake Drive, Itasca, Illinois 60143-2079.

Usos típicos y características especiales

En las aplicaciones típicas de estas pruebas de aprovechamiento de aplicación individual, existe un interés por diagnosticar discrepancias entre los diversos niveles de aprovechamiento o entre capacidad mental y aprovechamiento. De ahí que estas pruebas, por lo general, suponen la aplicación conjunta con alguna medida de capacidad mental. Determinar problemas de aprendizaje específicos es un objetivo frecuente de estos análisis. Debido a los énfasis especiales de las pruebas de esta categoría, a veces se denominan **baterías psicoeducativas**. Muchas aplicaciones de estas pruebas están estrechamente ligadas a leyes federales y estatales en relación con la identificación y tratamiento de varios problemas. Examinaremos estas leyes en el capítulo 16.

Evidentemente, la característica más distintiva de las pruebas de esta categoría es que se aplican de manera individual. Quizá más importante, su propósito es algo diferente de las pruebas de aprovechamiento de aplicación grupal, al menos, de tres maneras. Primero, mientras que los resultados promediados grupales son importantes para casi todas las pruebas de aplicación grupal, los resultados grupales son, en gran parte, irrelevantes para las pruebas de aplicación individual, excepto tal vez en algunos proyectos de investigación. Segundo, el foco de atención de las pruebas de aplicación grupal es a menudo el plan o programa de estudios en la misma medida en que el examinado individual. De hecho, como hemos señalado, las puntuaciones individuales pueden no ser siquiera informadas en algunas circunstancias cuando se emplean pruebas de aplicación grupal. La evaluación curricular no es un tema de las pruebas de aplicación individual. Por último, las pruebas individuales de aprovechamiento se centran detenidamente en el análisis de diferencias intraindividuales en el desempeño en distintas áreas de aprovechamiento y entre capacidad mental y aprovechamiento. Este tipo de análisis se realiza con algunas pruebas de aplicación grupal, pero incluso cuando esto ocurre, por lo general es menos crucial que en el caso de las pruebas de aplicación individual. (Una excepción a esta generalización es el *Wide Range Achievement Test* [Prueba de Aprovechamiento de Rango Amplio], WRAT, prueba de aprovechamiento de aplicación individual usada de manera casi exclusiva como un dispositivo rápido de exploración. Su aplicación suele tomar sólo cerca de 30 minutos y produce cinco puntuaciones.)

Aquí hay un problema especial que se debe considerar en el análisis de estas pruebas de aprovechamiento de aplicación individual. Como señalamos, hacen hincapié en las discrepancias entre capacidad y aprovechamiento y entre las puntuaciones de los perfiles. Los manuales presentan información acerca de los niveles de las diferencias requeridos para alcanzar significancia estadística. (Recordemos el concepto de error estándar de las diferencias que vimos en el capítulo 4.) Eso está bien; sin embargo, estos niveles de significancia se aplican a cada comparación de manera individual, no al conjunto entero de diferencias. Las probabilidades (de encontrar diferencias significativas) aumentan dentro del conjunto entero de diferencias, como es bien sabido en estadística. Si se

trabaja con una diferencia en el "nivel de significancia de .05", las probabilidades de encontrar *al menos una* diferencia significativa ascienden a casi 50% haciendo 15 comparaciones y a casi 80% haciendo 30. En otras palabras, si hacemos comparaciones suficientes, es casi seguro que encontraremos algo "significativo". El problema no es propio de las pruebas de aprovechamiento de aplicación individual, pero tiene una pertinencia particular en éstas debido al gran número de puntuaciones que generan y su énfasis en las comparaciones de puntuaciones.

Resumen de puntos clave 11-2

Algunas preguntas inquietantes acerca de las pruebas de aprovechamiento

- ¿Qué tan satisfactoria es la validez de contenido?
- ¿Qué tanto sentido tiene hablar de un continuo capacidad-aprovechamiento?
- ¿Qué hay con la motivación del examinado?
- ¿Cómo resolver el conflicto entre tiempo de evaluación e información diagnóstica?
- ¿Los reactivos de respuesta construida y de respuesta elegida miden de manera equivalente?

Algunas preguntas inquietantes acerca de las pruebas de aprovechamiento

Como señalamos al inicio de este capítulo, las pruebas de aprovechamiento, sin duda, son las que más se usan de todas las pruebas estandarizadas. Han tenido un papel importante en el desarrollo histórico del campo entero de las pruebas, han sido objeto de una enorme cantidad de investigación psicométrica y han evolucionado como resultado de dicha investigación. Ésta ha iluminado el uso de estas pruebas y guiado los progresos en su construcción. No obstante, quedan algunas preguntas inquietantes a las que dirigiremos nuestra atención aquí sin pretender dar respuestas definitivas.

Primero, hay algo que no es por completo satisfactorio en la noción de validez de contenido. Para estar seguros, existe un fuerte consenso de que la validez de contenido es el método clave para demostrar la validez de una prueba de aprovechamiento. Sin embargo, esta demostración siempre es fundamentalmente un juicio que no es susceptible de validación empírica. Los desarrollos más útiles dentro de la psicometría han requerido demostración empírica. La validez de contenido a veces parece como un callejón sin salida; si se cuestiona que una prueba tiene una buena o pobre validez de contenido, ¿hasta dónde llegamos de este modo? ¿No hay otra manera de aproximarnos al tema de la validez de la prueba? Puede ser que no podamos hacer algo mejor que estos juicios acerca de la validez de contenido. El problema de proporcionar evidencia adicional respecto de la validez se exacerba por las frecuentes revisiones del contenido de las pruebas. Ediciones nuevas o revisadas de muchas pruebas que vimos en este capítulo aparecen con frecuencia. Desde luego, es una meta admirable mantenerse actualizadas, pero esto también significa que es muy difícil obtener información de largo plazo.

El segundo tema se relaciona con la distinción entre capacidad y aprovechamiento introducida al principio del capítulo. La formulación que presentamos ahí parece tener sentido; sin embargo, no la aceptemos con demasiada facilidad. Primero, señalamos que muchos reactivos podrían aparecer con fundamentos igual de buenos en una prueba de aprovechamiento que de capacidad. Segundo, señalamos la correlación sumamente alta entre las pruebas de aprovechamiento y las de capacidad. En los análisis factoriales de estos dos tipos de pruebas, sentimos una fuerte presión para encontrar cualquier diferencia factorial entre ellas. ¿En verdad existe una diferencia entre estos constructos: capacidad y aprovechamiento? Desde luego, ésta no es una pregunta exclusivamente acerca de las pruebas de aprovechamiento, sino también de las de capacidad.

Tercero, está la pregunta acerca de la motivación del examinado. Ésta no es una pregunta seria cuando el examinado tiene algo en juego como un examen de licencia o una prueba de aprovechamiento que será usada para otorgarle un grado. En la prueba de aplicación individual, el examinador tiene la oportunidad de juzgar el nivel motivacional del examinado, pero ¿qué hay con el programa típico de evaluación escolar, una prueba ordenada por el estado o una evaluación de NAEP? Suponemos que los estudiantes

hacen un esfuerzo razonable en la prueba, pero en muchos casos esto parece una suposición frágil, sobre todo en la secundaria y el bachillerato. Esto puede no ser un problema, pero, al parecer, no sabemos mucho al respecto.

Cuarto, hay un conflicto aparentemente eterno entre el deseo de los usuarios de minimizar el tiempo de aplicación y maximizar la información diagnóstica. A los usuarios les gustaría tener pruebas de 15 reactivos que les dijeran todo acerca de las fortalezas y debilidades del estudiante.

Desafortunadamente, estos deseos son, en gran parte, incompatibles. Si queremos pruebas cortas, no tendremos mucha información diagnóstica. Si queremos información diagnóstica, necesitaremos pruebas largas. La aplicación adaptada para computadora computadora ayuda a mitigar un poco el conflicto, pero en esencia sigue rondándonos.

Por último, está la pregunta acerca de la diferencia entre lo que mide un reactivo de respuesta construida y uno de respuesta elegida. Está muy extendida la creencia de que el primero mide algo diferente, en especial algo más profundo y significativo, aunque la mayoría de las pruebas siguen basándose primordialmente en los reactivos de respuesta elegida. Incluso los expertos en psicometría sofisticada a veces son inconsistentes en este tema. Ellos piden, por un lado, enriquecer y mejorar las pruebas con reactivos de respuesta construida, pero, por el otro, combinan alegremente los resultados con reactivos de opción múltiple empleando procedimientos que suponen que todos los reactivos miden el mismo constructo. Este mismo tema surge en otras áreas de la evaluación, pero parece de particular agudeza en el campo de las pruebas de aprovechamiento. Quizá en los siguientes años de investigación habrá un consenso en esta cuestión. Quizá no.

Resumen

1. Las pruebas de aprovechamiento son, por mucho, las pruebas estandarizadas más usadas. Identificamos seis categorías importantes de ellas.
2. Los psicólogos tienen un papel importante en el mundo de las pruebas de aprovechamiento. Aportan su dominio de la elaboración de pruebas. Varias subespecialidades trabajan en las escuelas, donde las pruebas de aprovechamiento destacan.
3. La mayoría de las escuelas primarias y secundarias tienen programas de evaluación que incorporan varias pruebas con propósitos diferentes.
4. El movimiento de responsabilidad y el método basado en estándares de la educación han tenido una influencia importante en la manera en que se elaboran y utilizan las pruebas de aprovechamiento.
5. Las baterías de pruebas de aprovechamiento usadas en las escuelas primarias y secundarias son sistemas sofisticados y complejos que abarcan muchas pruebas y cubren muchos grados.
6. Las pruebas de aprovechamiento de área única se usan, en su mayor parte, para las evaluaciones de fin de cursos o fin de programas.
7. Los exámenes de licencia y certificación suelen ser pruebas largas y muy seguras que se usan para documentar la competencia.
8. Los procedimientos con referencia a una norma o a un criterio han evolucionado para establecer puntuaciones de corte en las pruebas de aprovechamiento.
9. Los estados han exigido programas de evaluación del aprovechamiento en escuelas primarias y secundarias. En la actualidad, los programas siguen los dictados de la ley NCLB.
10. Los programas nacionales e internacionales se concentran en los resúmenes grupales del desempeño de muestras elegidas cuidadosamente.
11. Las pruebas de aprovechamiento de aplicación individual, a menudo aplicadas junto con pruebas de capacidad, ayudan a detectar discrepancias entre capacidad y aprovechamiento o entre áreas de aprovechamiento.
12. Identificamos varias preguntas recurrentes acerca de la elaboración, uso e interpretación de las pruebas de aprovechamiento.

Palabras clave

anteproyecto de la prueba
batería
batería psicoeducativa
certificación
continuo capacidad-aprovechamiento
educación basada en estándares
ESEA
evaluación adaptada para computadora
inteligencia cristalizada
inteligencia fluida
licencia
NAEP
NCLB
NOCTI
pruebas de gran importancia
puntuación de corte
responsabilidad
TIMSS

Ejercicios

1. Piensa en tu propia historia escolar. ¿Qué pruebas estandarizadas de aprovechamiento recuerdas haber contestado? ¿Qué recuerdas de esas pruebas?
2. Entra a la página de internet de una de las editoriales que aparecen en el cuadro 11-3. ¿En qué hace hincapié la editorial acerca de la batería de aprovechamiento?
3. Para asegurarnos de que puedes usar la información del cuadro 11-4, responde estas preguntas:

- ¿Cuáles son los nombres de las pruebas de matemáticas del nivel P3?
- ¿Cuántos reactivos hay en la Batería Completa Intermedia 2 (I2)?
- ¿Cuánto tiempo se lleva la aplicación de la prueba Avanzado 1 (A1) Ciencia?
- ¿Qué nivel de la prueba se recomienda para el inicio del grado 2?

4. Para asegurarnos de que puedes usar la información de la figura 11-2, responde estas preguntas:

- ¿Cuál fue la puntuación estanina del estudiante en Vocabulario?
- ¿Cuál fue el equivalente de grado del alumno en Comprensión de lectura?
- ¿Cuántos reactivos (No. de reactivos) hubo en la prueba de Deletreo?

5. En la página de ETS Test Collection (http://www.ets.org/test_link/find_tests/), encuentra ejemplos de *pruebas de aprovechamiento de área única*. Escribe el nombre de un tema (p. ej., historia o ruso [history, Russian]) para ver qué clase de pruebas hay "allá afuera".
6. Entra al sitio nocti.org, elige un campo ocupacional de tu interés y encuentra el esquema de contenido de la prueba en esa área.
7. Entra a la página de internet del programa de evaluación de tu estado. Escribe [nombre del estado] testing program" en cualquier buscador de internet. Encuentra los resultados de una de las escuelas públicas de bachillerato de tu localidad. Describe brevemente los tipos de resúmenes que se presentan en el sitio.
8. Entra en la página de NAEP (http://nces.ed.gov/nationsreportcard/). ¿Qué tendencias se han informado con el paso del tiempo? ¿Cuántos años abarcan las tendencias que muestra NAEP?
9. En un sistema escolar conocido para ti, como al que asististe o donde trabaja algún familiar, consulta con el director de la escuela o el psicólogo escolar y construye un cuadro como el 11-2. Escribe los nombres exactos de las pruebas que se usan en la escuela.

Notas

[1] *Nota*: en el caso de los objetivos 4-7, puede ser útil construir una matriz como ésta:

Categoría	Ejemplos	Usos típicos	Características comunes

CAPÍTULO 12
Pruebas objetivas de personalidad

Objetivos

1. Definir el significado de una prueba objetiva de personalidad.
2. Identificar los usos típicos de las pruebas de personalidad.
3. Describir los problemas de la dirección y el falseamiento de respuestas y los métodos para tratar con estos problemas.
4. Describir los cuatro métodos de la elaboración de pruebas de personalidad y sus respectivas ventajas y desventajas.
5. Comparar las características comunes de los inventarios de personalidad integrales y de dominio específico.
6. Describir las principales características de al menos dos inventarios de personalidad.

Introducción

Empezaremos nuestro tratamiento del vasto campo de las pruebas de personalidad, intereses y actitudes considerando las pruebas objetivas de personalidad, de las cuales hay una cantidad impresionante. Será de utilidad empezar con algunas definiciones y clasificaciones de las pruebas que corresponden a esta categoría. A manera de anticipo, señalamos que las últimas partes de este capítulo se ocupan de pruebas relacionadas primordialmente con rasgos normales de personalidad, como extroversión y autoconcepto. El siguiente capítulo (13) se ocupa de los padecimientos anormales, patológicos o incapacitantes, como depresión y paranoia. Esta separación puede ser de utilidad, pero no es irrefutable. Mucho de lo que se explica en las siguientes secciones se aplica a las pruebas de ambas categorías, por lo que sirve de introducción a éste y el siguiente capítulo.

En las referencias a las **pruebas objetivas de personalidad**, la palabra "objetivas" tiene un significado muy especial: que los reactivos se pueden calificar de manera objetiva sin necesidad de emplear el juicio, por lo que no se requiere un entrenamiento profesional para ello. Un oficinista o una máquina pueden calificar una prueba objetiva de personalidad. En la práctica, esto significa que las respuestas son de opción múltiple. Recordemos del capítulo 6 que un término más técnico para referirnos a esos reactivos es el de respuesta cerrada. El examinado elige una respuesta de un conjunto fijo de alternativas. El cuadro 12-1 muestra ejemplos de los formatos típicos de los reactivos de pruebas objetivas de personalidad. Sólo se trata de ejemplos, no de una lista exhaustiva de formatos de respuesta en este tipo de pruebas.

Cuadro 12-1. Formatos de respuesta comunes en las pruebas objetivas de personalidad

Verdadero–Falso
De acuerdo–En desacuerdo
Verdadero–Falso–No sé
Totalmente de acuerdo–De acuerdo–No sé/Neutral–En desacuerdo–Totalmente en desacuerdo
Elegir entre dos afirmaciones, por ejemplo: Marque lo que mejor lo describe a usted:
A. Me agradan la mayoría de las personas. B. Suelo trabajar muy duro.

Aunque la naturaleza del formato de respuesta es la característica que define las pruebas objetivas de personalidad, también existe otra característica: la naturaleza del tronco del reactivo. Por lo común, el tronco del reactivo consiste en afirmaciones sencillas de una sola oración, de las cuales se muestran ejemplos en el cuadro 12-2; sin embargo, algunos reactivos pueden ser de sólo una palabra. Por ejemplo, puede haber una lista de adjetivos (amigable, puntual, perezoso, etc.) y el examinado marca los que mejor lo describan. Es muy poco común que el tronco del reactivo implique más de una oración en las pruebas objetivas de personalidad.

Cuadro 12-2. Tipos de afirmaciones que pueden aparecer en pruebas objetivas de personalidad

1. Me agradan la mayoría de las personas.
2. Rara vez pierdo los estribos.
3. La mayoría de las personas trabaja duro.
4. A veces escucho voces.
5. Aquí, el clima es espantoso.
6. Me pongo furioso con facilidad.
7. La mayoría de los niños son malos.
8. Los periódicos están llenos de mentiras.
9. No se puede contar con la mayoría de las personas para un trabajo honesto.
10. La vida es un costal de problemas.
11. En mi cabeza suceden cosas extrañas.
12. Tengo amigos cercanos en la mayoría de ciudades.
13. Suelo terminar las cosas a tiempo.

Hacemos el contraste entre pruebas objetivas de personalidad y medidas *proyectivas* de la personalidad. (Lo esperable sería contrastar pruebas *objetivas* con pruebas "subjetivas", pero éste no es el caso.) En el capítulo 14, nos ocuparemos de las pruebas proyectivas, las cuales emplean un formato de respuesta abierta o libre, por lo que se requiere del juicio para calificarse. Así, el método de calificación es la distinción crucial entre pruebas objetivas y proyectivas, aunque los troncos de los reactivos también suelen ser diferentes.

En este campo, encontramos varios términos más. Primero, a menudo se denomina **inventario** a una prueba objetiva de personalidad; encontramos esta extraña designación sólo tratándose de pruebas de personalidad y de intereses vocacionales. Este término simplemente quiere decir prueba. De manera consistente con la práctica en este campo, usamos los términos prueba e inventario como sinónimos en este capítulo. Segundo, señalamos que algunas fuentes usan el término **prueba estructurada** en vez de objetiva para caracterizar las pruebas que tratamos en este capítulo. Preferimos el término objetiva en lugar de estructurada, pero no vale la pena hacer objeciones de poca monta acerca de cuál es mejor.

Usos de las pruebas objetivas de personalidad

Existen cuatro usos primarios de las pruebas objetivas de personalidad (véase Resumen de puntos clave). *P*rimero y más importante, los psicólogos clínicos las usan para realizar una evaluación estandarizada de los rasgos y características de personalidad. Este uso motivó la elaboración de muchas pruebas de personalidad, entre las que se encuentran las más usadas. Por ejemplo, es común usar una o más de estas pruebas al inicio de la evaluación de un cliente. Una de las pruebas también puede usarse después en el proceso de tratamiento para documentar los cambios en el individuo. Desde luego, los psicólogos clínicos también obtienen información de otras fuentes, como entrevistas, expedientes médicos, escolares y laborales, entrevistas con miembros de la familia y otros tipos de

pruebas, sobre todo las de capacidad mental.

Resumen de puntos clave 12-1

Principales usos de las pruebas objetivas de personalidad
Clínico
Asesoría
Selección de personal
Investigación

Podemos incluir el uso **forense** como una subcategoría del clínico, que se refiere a cualquier uso en procesos legales; por ejemplo, un juez puede solicitar una evaluación de la personalidad de un acusado en un caso penal o de los padres en un pleito por la custodia de un niño. En el capítulo 16, pp. 423-424a», se describe con más detalles el uso forense. Otra subcategoría del uso clínico es la evaluación neuropsicológica, que tratamos en el capítulo 10. Recordemos que, aunque se centra en las funciones cognitivas, la evaluación neuropsicológica, por lo general, incluye la personalidad casi siempre con ayuda de un inventario objetivo.

El segundo uso de las pruebas objetivas de personalidad es para propósitos de orientación. Por ejemplo, cuando se trabaja con una pareja en un contexto de orientación marital, puede ser útil tener perfiles de los cónyuges. Cuando se trabaja con un estudiante universitario que se encuentra bajo un estrés inusual (pero no patológico), el orientador puede emplear un inventario de personalidad.

Tercero, las pruebas de personalidad se usan a veces para la selección de personal. Desde luego, las pruebas de capacidad y de aprovechamiento se emplean con mucha frecuencia para este propósito; sin embargo, en el caso de algunos puestos, también se usan las de personalidad. El punto central puede ser identificar individuos con características de personalidad que predigan el éxito en cierta ocupación o identificar individuos que pueden ser empleados "problema", como indican, por ejemplo, las que a menudo se denominan "pruebas de integridad".

Cuarto, las pruebas objetivas de personalidad se usan mucho en la investigación sobre la personalidad humana. De hecho, muchas de ellas se elaboraron primordialmente para investigar constructos de personalidad, más que su uso aplicado. El uso de estas pruebas en la investigación se divide en tres amplias categorías. Primero, se usan en la investigación básica sobre la estructura misma de la personalidad, lo cual va de la mano con el desarrollo de las teorías de la personalidad. Segundo, hay una enorme cantidad de investigación relacionada con la aplicación clínica de estas pruebas, su validez y confiabilidad en diferentes tipos de poblaciones, por ejemplo, ¿cómo se desempeñan los individuos con depresión o abuso de sustancias en una prueba específica? Tercero, las pruebas de personalidad se emplean en diversos estudios para determinar la manera en que las características de personalidad se relacionan con otras variables, por ejemplo, ¿las variables de personalidad ayudan a predecir el éxito en la escuela o se relacionan con la

capacidad para aprender sílabas sin sentido? ¿Cuáles son las características de personalidad de los artistas más creativos o los empresarios de la tecnología de punta? Investigar las correlaciones de la personalidad con diversas variables es un tema que provoca una fascinación sin límites.

Clasificación funcional de las pruebas objetivas de personalidad

Para ayudarnos a tratar con la asombrosa cantidad de pruebas objetivas de personalidad, será útil introducir un sistema de clasificación funcional, que hace hincapié en las distinciones que surgen en la práctica, más que una clasificación de gran elegancia teórica. El cuadro 12-3 muestra un sistema de clasificación bidimensional; en cada categoría de este cuadro de 2 × 2, la figura incluye ejemplos de pruebas, algunas de las cuales se describen con mayor detalle en este capítulo.

Cuadro 12-3. Clasificación funcional de las pruebas objetivas de personalidad[a]

Orientación	**Ámbito de cobertura**	
	Integral	**Dominio específico**
Normal	*Edwards Personal Preference Schedule* *Sixteen Personality Factor InventoryNEO* *Personality Inventory*	*Piers-Harris Children´s Self-Concept Scale* *Rotter Locus of Control Scale* *Bem Sex Role Inventory*
Anormal	*MMPI* *Millon Clinical Multiaxial Inventory* *Personality Assessment Inventory*	*Beck Depression Inventory* *State-Trait Anxiety Inventory* *Suicidal Ideation Questionnaire*

[a] Gracias a John C. Norcross por sugerir este esquema de clasificación.

En el lado izquierdo del cuadro 12-3, distinguimos entre pruebas que se orientan hacia la personalidad normal o anormal. Algunas pruebas objetivas de personalidad están diseñadas primordialmente para abordar características anormales, patologías y problemas, que constituyen el repertorio de los psicólogos clínicos. Las pruebas de esta categoría representan a menudo el nombre de la anormalidad que se mide (p. ej., ansiedad, depresión, etc.) en el título de la prueba o de las escalas dentro de ésta. Otras pruebas están diseñadas para medir rasgos de personalidad en la *población normal*; estas pruebas pueden ser usadas por psicólogos consejeros y en la investigación de una amplia variedad de temas relacionados con variables de personalidad de la vida cotidiana. Los títulos de estas pruebas, por lo general, no son descriptivos, mientras que los nombres de las escalas que forman parte de ellas hacen referencia a rasgos o constructos típicos de la personalidad, por ejemplo, extroversión o autoconcepto. Esta distinción entre lo normal y lo anormal es justo la base para separar este capítulo del siguiente, sobre todo en los ejemplos de las pruebas que revisamos.

En la parte alta del cuadro 12-3, distinguimos entre las pruebas que intentan brindar una cobertura muy amplia y aquellas cuyo objetivo es más estrecho; llamamos

inventarios integrales a las primeras y de **dominio específico** a las segundas. Vale la pena seguir esta distinción con mayor detalle.

¡Inténtalo!

Clasifica las siguientes pruebas en el sistema que se presenta en el cuadro 12-3.
Escala de Autoconcepto de Tennessee
Inventario Básico de Personalidad
Prueba de Chequeo de Abuso de Drogas

Cuadro 12-4. Ejemplos de pruebas objetivas integrales de personalidad

Prueba	Editorial	Reactivos
Edwards Personal Preference Schedule (EPPS)	*Psychological Corporation*	225
California Psychological Inventory (CPI)	*CPP*[a]	434
Minnesota Multiphasic Personality Inventory-2 (MMPI-2)	*University of Minnesota*	567
Myers-Briggs Type Indicator (MBTI)	*CPP*	93-290[c]
NEO PI-3	*PAR*[b]	240
Sixteen Personality Factor Inventory (16PF, 5th ed.)	*Institute for Personality and Ability Testing*	185
Million Clinical Multiaxial Inventory-III (MCMI-III)	*Pearson*	175
Personality Research Form (PRF)	*Research Psychologists Press*	300-440[d]
Personality Assessment Inventory	*PAR*	344

[a] Antes *Consulting Psychologists Press.*
[b] Antes *Psychological Assessment Resources.*
[c] Las diferentes formas del MBTI varían de 93 a 260 reactivos.
[d] Existen seis formas del PRF con números de reactivos que van de 300 a 440.

Inventarios integrales: características en común

El cuadro 12-4 enumera algunos de los inventarios integrales de personalidad más usados. La revisión de las entradas típicas de esta categoría revela que comparten ciertas características en común, más allá del hecho de que intentan ser relativamente integrales y usar un formato de respuesta cerrada. Antes de examinar algunos de estos inventarios en detalle, será útil identificar estas características; después, revisaremos las características en común de las entradas en la categoría de dominio específico. Desde luego, dentro de cada categoría, observaremos algunas excepciones a la lista de características en común, aunque ésta no deja de ser útil en modo alguno.

1. Los inventarios integrales tienden a tener una gran cantidad de reactivos, por lo

general de 200 a 600. ¡Son muchos reactivos! En el cuadro 12-4 se puede observar el número de reactivos de algunos de los inventarios integrales más populares.

2. Debido a que los troncos de los reactivos son cortos y las respuestas son de opción múltiple, los examinados, por lo común, contestan estos inventarios en *30 o 60 min* a pesar del gran número de reactivos, aunque algunos de los inventarios más largos pueden requerir hasta 90 min. En contraste con las pruebas de capacidad o aprovechamiento, los inventarios de personalidad no tienen límites de tiempo. Sin embargo, se suele animar al examinado a responder con rapidez los reactivos en vez de trabajar detenidamente en ellos.

3. Las pruebas de esta categoría tienden a producir muchas puntuaciones. Una entrada típica informa de 15 a 20 puntuaciones. Incluso cuando una de estas pruebas se concentra en un número específico de rasgos, a menudo produce puntuaciones adicionales; por ejemplo, el *NEO PI* se considera uno de los mejores ejemplos de una medida de los "Cinco Grandes" rasgos de personalidad. De hecho, ofrece puntuaciones de cada uno de estos rasgos, pero también produce 30 puntuaciones de "facetas", lo que da un total de 35 puntuaciones. El 16PF, como su nombre lo indica, produce puntuaciones de 16 rasgos, pero también cinco puntuaciones basadas en la combinación de los 16 rasgos, que corresponden más o menos a los Cinco Grandes rasgos de personalidad, además de tres puntuaciones de tendencias de respuesta, lo cual da un gran total de 24 puntuaciones.

En algunos casos, es difícil decir con exactitud cuántas puntuaciones producen estos inventarios porque, además de las puntuaciones "regulares" previstas desde su origen, estudios subsiguientes han identificado otros subconjuntos de reactivos que producen nuevas puntuaciones. El ejemplo clásico de este fenómeno es el MMPI-2, pues los investigadores siguen encontrando reactivos que diferencian un grupo de personas de otro, por lo que combinan estos reactivos para formar una nueva escala.

La multiplicidad de puntuaciones de los inventarios integrales da origen a una importante consideración que afectará nuestra evaluación de las pruebas. Aunque el inventario completo tiene muchos reactivos, como señalamos, las escalas individuales a menudo no tienen tantos, lo cual limita su confiabilidad. A veces leemos que las pruebas de personalidad no son tan confiables como las de capacidad y aprovechamiento, pero, al menos en algunos casos, esta generalización surge porque las escalas de personalidad se basan en relativamente pocos reactivos.

En el campo de las pruebas de capacidad y aprovechamiento, es común informar una puntuación total además de varias subpuntuaciones. Es interesante señalar que esto no sucede en el campo de la personalidad, pues ningún inventario combina todas sus subpuntuaciones en una sola puntuación total. Según parece, los psicólogos no creen en la existencia de un factor "*g*" de la personalidad.

4. La cuarta característica de los inventarios integrales es que tienen muchas aplicaciones. Muchos tipos de psicólogos las usan en diversos contextos y con distintos propósitos. También se usan en una enorme cantidad de proyectos de investigación para mejorar nuestra comprensión de la personalidad humana e identificar diferencias

grupales en padecimientos clínicos.

5. Los inventarios integrales elaborados en años recientes, por lo general, hacen un esfuerzo deliberado para ofrecer grupos de normas representativos a nivel nacional bien definidos. Esto no ocurrió con las primeras versiones de las pruebas de esta categoría (ni con algunos inventarios integrales que no se han revisado en años recientes). Aunque las pruebas se usaron en un contexto nacional, sus normas no son representativas, sino notablemente locales. Las mejoras a los inventarios integrales en este aspecto podrían contarse entre las contribuciones importantes de la tradición psicométrica para llevar a cabo revisiones de las pruebas. A menudo, lamentamos la falta de progreso que resulta del proceso de revisión, pues mejores normas de los inventarios integrales de personalidad ofrecerían, al menos, más luz al escenario. Una visión extrema atribuiría el cambio a las presiones competitivas comerciales más que a la sensibilidad ante las críticas profesionales.

6. Una última característica emergente de los inventarios integrales, al menos en el caso de las ediciones recientes, es la provisión de informes interpretativos. Pruebas como MMPI-2, NEO PI-3, 16 PF y Millon tienen narrativas extensas que incorporan interpretaciones basadas en estudios de validez, así como en comparaciones normativas. La producción de estos informes interpretativos ahora está estrechamente vinculada a la transmisión electrónica de respuestas. Un examinado puede introducir las respuestas al inventario de personalidad en línea en la computadora de la editorial y, momentos después, el clínico puede tener a su disposición una interpretación narrativa extensa del perfil de puntuaciones del examinado. Éste es un desarrollo importante reciente en el campo de las pruebas psicológicas.

¡Inténtalo!

Para ver los diversos proyectos de investigación que emplean inventarios integrales de personalidad, escribe el nombre de cualquiera de los que aparecen en el cuadro 12-4 como palabra clave en cualquier base de datos de tu biblioteca. PyschINFO sería ideal. Para no obtener un número abrumador de referencias, limita la búsqueda a un periodo de cinco años.

Cuadro 12-5. Ejemplos de pruebas objetivas de personalidad de dominio específico

Prueba	**Editorial**	**Reactivos**
Beck Depression Inventory-II	Pearson	21
Beck Scale for Suicide Ideation	Pearson	21
Bem Sex Role Inventory	Mind Garden	60[c]
Children's Depression Inventory-2	MHS[a]	28[c]
Coopersmith Self Esteem Inventory	MHS	58[c]
Eating Disorder Inventory-3	PAR	91
Piers-Harris Children's Self-Concept Scale-2	PAR	60

State-Trait Anxiety Inventory	Mind Garden	40
Suicidal Ideation Questionnaire	PAR	30
Tennessee Self-Concept Scale	WPS[b]	(Adulto) 82[c] (Niño) 76[c]

[a] Antes *Multi-Health Systems*, Inc.
[b] Antes *Western Psychological Services*, Inc.
[c] Cada una también tiene una forma corta con menos reactivos, por lo general alrededor de 20.

Pruebas de dominio específico: características en común

El cuadro 12-5 muestra ejemplos de algunas de las numerosas **pruebas de dominio específico**. En general, las características de estas pruebas son las opuestas a las de las pruebas integrales. Sin embargo, hay algunas excepciones a esta generalización. Las pruebas de dominio específico manifiestan las siguientes características.

1. Tienen relativamente pocos reactivos; no es raro que estas pruebas tengan menos de 30 reactivos, pero sí lo es que tengan 100. A pesar de su brevedad, a menudo tienen una excelente confiabilidad de consistencia interna, justo porque se centran en un dominio restringido.

2. Debido a su reducido número de reactivos, estas pruebas pueden contestarse de manera rápida, pues requieren sólo de 10 a 15 min.

3. Las pruebas de dominio específico, por lo general, tienen pocas puntuaciones, a menudo sólo una. Cuando tienen más de una puntuación, las diversas puntuaciones tienen una relación conceptual estrecha. Por ejemplo, puede haber puntuaciones separadas para distintas facetas del autoconcepto o varias áreas de la depresión. Mientras que las puntuaciones diferentes de las pruebas integrales nunca se suman para obtener una puntuación total, las puntuaciones diferentes de una prueba de dominio específico, debido a que pueden relacionarse de manera estrecha, pueden sumarse para obtener una puntuación total.

4. Las pruebas de dominio específico no tienen un amplio rango de aplicaciones. Por lo general, tienen usos y audiencias muy definidos. Las medidas de autoconcepto pueden ser una excepción a esta observación.

5. Por lo general, las pruebas de dominio específico tienen grupos de estandarización muy limitados. En algunos casos, no se presentan normas; en otros, los grupos de estandarización son casi ejemplos puros de "normas por conveniencia" (véase en el capítulo 6 una discusión sobre este tipo de grupos de estandarización). Es decir, las normas se basan en uno o más grupos disponibles, pero es casi nulo el intento de referir los grupos a alguna población general.

6. Por último, la calificación y los informes tienden a ser muy sencillos en el caso de las pruebas de dominio específico. No son difíciles de calificar y, debido a que tienen pocas puntuaciones, no se necesitan esquemas interpretativos complicados.

Antes de dejar la distinción entre pruebas integrales y de dominio específico, debemos

comentar las circunstancias de su aplicación práctica. Los inventarios integrales tienden a usarse cuando el psicólogo necesita investigar muchas posibilidades en un individuo; cuando sabemos qué problema tiene un individuo o necesitamos estudiar muchas facetas de su personalidad, es apropiado usar una batería integral. Sin embargo, si tenemos una idea muy clara de en qué debemos centrarnos, entonces es más apropiada una prueba de dominio específico. Lo que se pone en la balanza es la amplitud de la información y el tiempo (y los costos). Una prueba integral proporciona más información, pero requiere más tiempo, mientras que una de dominio específico es corta y sencilla, pero ofrece información más limitada. (Véase cuadro 12-6.)

Cuadro 12-6. Resumen de las principales diferencias entre pruebas de personalidad integrales y de dominio específico

Característica	Integrales	De dominio específico
Número de reactivos	Muchos; por lo general, varios cientos	Pocos; por lo general, de 20 a 80
Tiempo de aplicación	45 min en promedio, pueden ser más	Por lo general, de 10 a 15 min
Número de puntuaciones	Muchas	Pocas, a menudo sólo una
Rango de aplicaciones	Amplio	Restringido
Grupos de estandarización	Buena representatividad[a]	Muy limitados
Informes	Elaborados; a menudo narrativos[a]	Simples

[a] Esta descripción sólo se aplica a los inventarios integrales recientemente creados o revisados.

Problemas especiales de la dirección y el falseamiento de respuestas «315-317a

Las pruebas objetivas de personalidad –tanto las que se enfocan en rasgos normales como las que buscan padecimientos anormales– están plagadas de problemas relacionados con la dirección y el falseamiento de respuestas. La **dirección de las respuestas** es la tendencia, consciente o inconsciente, de una persona a responder los reactivos en cierto modo, independientemente de los verdaderos sentimientos de ella. Otros términos que se encuentran con frecuencia en la literatura sobre este tema son **distorsión de respuestas** (los verdaderos sentimientos de una persona son distorsionados de cierta manera por la dirección de las respuestas) y **manejo de impresiones** (una persona trata de causar cierta impresión por medio de sus respuestas). Un concepto similar es **estilos de respuesta**; los más comunes son la tendencia a responder de modos socialmente deseables al responder estar de acuerdo o en desacuerdo con determinadas afirmaciones. Las **respuestas socialmente deseables**, por lo general, son las que la sociedad aprueba, por ejemplo, ser simpático y trabajador. La tendencia a estar de acuerdo, también conocida como aquiescencia o tendencia a decir sí, significa que la persona puede estar de acuerdo con casi cualquier afirmación, mientras que la tendencia a estar en desacuerdo (la tendencia a decir no) significa que la persona se inclina a estar

en desacuerdo con casi cualquier afirmación.

Estas tendencias, por sí mismas, podrían considerarse características de personalidad que merecen ser medidas. De hecho, algunas pruebas producen puntuaciones separadas de la dirección de respuestas; sin embargo, se crea un problema cuando interfieren en la medición de otros rasgos de personalidad. Consideremos el rasgo "cordialidad"; la gente varía en este rasgo desde muy cordial hasta muy hostil. Por lo general, consideramos que es socialmente deseable ser cordial. Si una persona responde un reactivo en la dirección de "cordial" porque esto es lo socialmente deseable a pesar de que, en realidad, la persona no es muy cordial, entonces tenemos un problema. Lo que queremos es que la persona responda en la dirección de "cordial" sólo si, en verdad, es cordial, independientemente del hecho de que esta respuesta sea socialmente deseable. El truco es desenredar la variable cordialidad y separarla de la variable deseabilidad social. Como veremos, las pruebas de personalidad han adoptado diversas estrategias para tratar con éstas y otras tendencias de respuesta.

El *falseamiento* es un intento deliberado para causar una impresión favorable o desfavorable. (Un término más sofisticado es *disimulación.*) El **falseamiento positivo** significa causar una impresión favorable y el **falseamiento negativo** significa causar una impresión desfavorable; a veces se le llama **fingimiento**. En muchas circunstancias, el examinado no tiene motivación para falsear las respuestas o puede estar muy satisfecho de presentar un retrato honesto de sí mismo. Sin embargo, en otras circunstancias, puede haber una considerable motivación para causar una impresión especialmente favorable o desfavorable. Por ejemplo, una persona que solicita un trabajo como vendedor puede aparentar ser extrovertida y ambiciosa, mientras que una persona acusada de asesinato puede representar una imagen de inestable, incluso delirante para que la defensa pueda reducir su culpabilidad. Desde luego, el solicitante para el trabajo de vendedor en verdad puede ser extrovertido y ambicioso, y el acusado de asesinato puede tener en realidad delirios. Al igual que con el tema de la deseabilidad social, el truco es desenredar el falseamiento de las características reales de personalidad.

Resumen de puntos clave 12-2

Cuatro estrategias principales para tratar la dirección y el falseamiento de respuestas

1. Examinar las respuestas a los reactivos con frecuencias empíricas extremas
2. Revisar la consistencia en reactivos iguales o similares
3. Balancear la dirección de los reactivos
4. Usar el método de elección forzada con los reactivos correspondientes a la variable pertinente.

Estrategias para tratar la dirección y el falseamiento de respuestas

Los psicólogos han inventado diversas estrategias para detectar y/o reducir las influencias de la dirección y el falseamiento de respuestas en las puntuaciones de las pruebas de personalidad. Es esencial comprender estas estrategias, ya que nos referiremos a ellas

con frecuencia en nuestro tratamiento de pruebas de dominio específico más adelante en este capítulo y las encontrarás cuando revises los manuales de este tipo de pruebas. La mayoría de las estrategias cae en una de cuatro categorías importantes (véase Resumen de puntos clave).

La primera categoría incluye referencias a **frecuencias empíricas extremas** de los grupos normales. Un grupo "normal" aquí significa una muestra representativa de la población que no tiene más que el número usual de individuos con desviaciones de personalidad; los miembros de este grupo no tienen una motivación particular para responder de una manera que no sea honesta. La frecuencia real de las respuestas a los reactivos se determina para dicho grupo. Puede resultar que prácticamente nadie en el grupo diga que "le gusta la mayoría de la gente" o que tiene "amigos cercanos en la mayoría de las ciudades". De ahí que, cuando un examinado marca "Verdadero" en uno de estos reactivos, sospechamos que "falsea positivamente" sus respuestas o que puede estar demasiado influido por la tendencia a la deseabilidad social. Puede resultar que prácticamente nadie, quizá incluso entre las personas perturbadas, indique que "los periódicos están llenos de mentiras". Una persona que marca este reactivo como "Verdadero" es sospechosa de "falsear negativamente" sus respuestas. Desde luego, las conclusiones acerca del falseamiento no se basan en las respuestas a sólo uno o dos reactivos, pues un inventario puede contener una docena o más de reactivos de este tipo, que producen una puntuación del presunto falseamiento.

La segunda estrategia determina la **consistencia de respuestas** en reactivos similares. En el cuadro 12-2, los reactivos 2 y 6, y 3 y 9 forman pares de reactivos similares, aunque la redacción es en el sentido opuesto dentro de cada par. Para estar seguros, hay ligeras diferencias en los matices del significado en los reactivos de cada par, pero las semejanzas en el significado son mucho mayores. Esperaríamos que la persona que marca "V" en el reactivo 2 marque "F" en el reactivo 6, y lo mismo en los reactivos 3 y 9. Además, podríamos demostrar de manera empírica una correlación muy alta en las respuestas de estos pares de reactivos. ¿Qué haríamos con un examinado que marcó "V" en los reactivos 2 y 6, y 3 y 9? Podría tratarse de alguien que dice sí a todo. Si todas las respuestas son "F", podría tratarse de alguien que dice no a todo. Si las respuestas son "V" en un par y "F" en otro, el examinado podría haber respondido al azar. Determinando la consistencia de las respuestas en, digamos, 15 pares de reactivos similares, podemos obtener una puntuación de consistencia. Podemos determinar que, en nuestro grupo normal, la mayoría de las personas es consistente en los 15 pares y muy poca es inconsistente en más de dos de estos pares. De ahí que las puntuaciones de tres o más pares pueden llevarnos a cuestionar la validez de otras puntuaciones del inventario.

¡Inténtalo!

¿Qué persona parece mostrar respuestas inconsistentes en estos dos reactivos?
Las respuestas son V = Verdadero y F = Falso.

	Persona

Reactivo	A	B	C
Me llevo bien con la mayoría de las personas	V	F	F
Me desagrada la mayoría de las personas que conozco	F	V	F

La tercera estrategia trata con las tendencias a responder sí o no a todo *equilibrando la dirección* de los reactivos. Supongamos que medimos el rasgo de "cordialidad". Podríamos empezar escribiendo reactivos como los primeros dos del cuadro 12-7. Una respuesta "Verdadero" a estos reactivos se encuentra en la dirección de la "cordialidad". Sin embargo, si todos los reactivos de esta escala tienen esta dirección, es decir, si la respuesta Verdadero indica cordialidad, *y* un examinado tiene la tendencia a decir sí a todo, entonces la medición de la cordialidad estaría enredada con esta tendencia. Para ayudar a controlar esto, incluimos reactivos como el 3 y 4 del cuadro 12-7. En el caso de estos reactivos, una respuesta "Falso" sería calificada en la dirección de la cordialidad. Así, la puntuación máxima de cordialidad resulta de las respuestas V, V, F y F en estos cuatro reactivos, respectivamente. No es difícil decir cómo se aplica el mismo razonamiento al control de la tendencia a decir no a todo.

Cuadro 12-7. Ejemplos del equilibro en la dirección de los reactivos

Reactivo	Respuestas en dirección de la cordialidad	
	Verdadero (V)	Falso (F)
1. Disfruto estar con amigos.	V	
2. A menudo visito lugares con amigos.	V	
3. Los amigos traen más problemas de lo que valen.		F
4. Tengo muy pocos amigos		F

Muchos inventarios de personalidad intentan tener cierto equilibro, aunque no siempre exacto, en la dirección de los reactivos de un rasgo. Un intento nada elegante busca este equilibro mediante la simple inserción de "no" en el tronco de algunos reactivos. A menudo, esto lleva a una situación embarazosa en la que el examinado debe emplear una doble negación para expresar un sentimiento positivo. Por ejemplo, en lo que respecta a los reactivos del cuadro 12-7, un reactivo como "No tengo muchos amigos" requiere de una respuesta "Falso" para indicar cordialidad, lo cual es una solución torpe.

La cuarta estrategia, creada principalmente para tratar con la variable de deseabilidad social, requiere que el examinado elija entre afirmaciones que corresponden a la deseabilidad social. Esta correspondencia se determina de manera empírica con jueces que valoran la deseabilidad social de la respuesta en cierta dirección por cada reactivo. Los reactivos con niveles similares de deseabilidad social se ubican en pares (o tríadas). Se pide al examinado que elija la afirmación que mejor lo describa. Ya que las afirmaciones en par son más o menos iguales en deseabilidad social, su elección debe estar determinada por su personalidad más que por la tendencia a dar una respuesta

socialmente deseable. La correspondencia puede ser con alguna variable diferente a la deseabilidad social, pero esta técnica suele emplearse para controlar esta variable.

Consideremos los pares de afirmaciones del cuadro 12-8. Supongamos que hemos determinado de manera empírica que los pares de afirmaciones tienen una correspondencia con la deseabilidad social. Supongamos, además, que la afirmación 1A se califica en una escala de "escrupulosidad", las afirmaciones 1B y 2A (con una calificación invertida) se califican en una escala de "sociabilidad" y 2B se califica en una escala de "control personal". Ya que los pares tienen una correspondencia con la deseabilidad social, las elecciones del examinado deben estar determinadas por su estatus en escrupulosidad, sociabilidad y control personal más que por una tendencia a dar respuestas socialmente deseables (o indeseables).

Cuadro 12-8. Elección entre afirmaciones correspondientes a la deseabilidad social

Instrucciones: De cada par, elige la afirmación (A o B) que *mejor* te describa.	
1A. Suelo trabajar duro.	1B. Me agrada la mayoría de las personas que conozco.
2A. A menudo, me siento intranquilo entre la gente.	2B. Con frecuencia, pierdo los estribos.

Después de determinar la puntuación de una tendencia de respuesta o falseamiento, disponemos de dos métodos principales para usar esta información. Primero, la puntuación de la tendencia de respuesta (p. ej., decir sí a todo) puede llevar a ajustes en las puntuaciones de los rasgos de personalidad que son de interés primario. Segundo, la puntuación de una tendencia de respuesta puede llevar a invalidar todas las otras puntuaciones o, al menos, a tener serias dudas acerca de su validez. Por ejemplo, si existen demasiadas respuestas inconsistentes, podemos suponer que el examinado respondió al azar, de manera descuidada o incoherente, de modo que ninguna de las puntuaciones es un indicador útil de los rasgos de personalidad.

Las puntuaciones de la consistencia y las diferentes tendencias de respuesta se denominan a menudo índices de validez de las pruebas objetivas de personalidad. El término es desafortunado, porque no significa lo mismo que en el caso de la validez de las pruebas (revisada en el capítulo 5). En general, la validez de las pruebas se refiere al grado en que la prueba mide lo que pretende medir, mientras que los índices de validez, en el caso de las pruebas de personalidad, sólo se refiere a si las respuestas del examinado son sospechosas.

¡Inténtalo!

La tarea consiste en valorar los reactivos del cuadro 12-2 en relación con la deseabilidad social. Usa una escala de 10 puntos, en la que 10 es lo más deseable y 1 lo menos deseable. ¿Cuáles son los dos reactivos que ubicarías cerca del 10? ¿Cuáles son los dos reactivos que ubicarías cerca del 1?
Ubicación cercana a 10:

Reactivos _____ y _____
Ubicación cercana a 1:
Reactivos _____ y _____
Compara tus valoraciones con las de algún compañero.

Resumen de puntos clave 12-3

Principales enfoques para elaborar pruebas objetivas de personalidad

- Contenido
- Clave del criterio
- Análisis factorial
- Teoría

Principales enfoques para elaborar pruebas de personalidad

Existen cuatro métodos principales para elaborar pruebas objetivas de personalidad (véase Resumen de puntos clave). Más que en otros tipos de pruebas, el método de elaboración ofrece un marco para comprender estas pruebas; de ahí que, antes de describir pruebas específicas de personalidad, bosquejaremos estos enfoques importantes para elaborar pruebas y describiremos sus ventajas y desventajas. Debemos tener en mente al principio que la mayoría de las pruebas actuales emplean alguna combinación de estos métodos.

Método de contenido

El método de contenido, también conocido como método lógico o racional, elabora los reactivos y las escalas de la prueba con base en una comprensión sencilla y directa de lo que queremos medir. Podría denominarse método de sentido común para elaborar pruebas. Por ejemplo, si queremos medir extroversión, entonces planteamos un conjunto de preguntas acerca de las relaciones con las personas. Si queremos saber si una persona tiene una tendencia hacia la hipocondría (preocupación excesiva por la salud personal), planteamos preguntas acerca del miedo a los gérmenes, pensamientos sobre enfermedades, entre otras. El método de contenido se aproxima a la forma escrita de lo que se podría cubrir en una entrevista.

Este método fue la base para elaborar la primera prueba de personalidad ampliamente utilizada, el *Woodworth Personal Data Sheet*. Como señalamos en el capítulo 1, esta prueba presenta en forma escrita las preguntas que un clínico podría plantear de manera

habitual en una entrevista individual. Sin embargo, al ponerla por escrito, la prueba puede aplicarse de manera grupal, lo cual ahorra tiempo y costos.

Ventajas y desventajas

El método de contenido tiene la obvia ventaja de la sencillez. Además, dada cierta comprensión razonable del constructo que se medirá, suele ser fácil generar reactivos empleando este método, que también tiene una buena validez aparente. El mayor inconveniente del método de contenido es que las respuestas pueden ser distorsionadas por los estilos de respuesta y los esfuerzos, conscientes o inconscientes, de falsear las respuestas de manera positiva o negativa. De hecho, el reconocimiento de esta desventaja llevó a elaborar otros métodos para construir pruebas de personalidad.

En la práctica contemporánea, no se utiliza únicamente el método de contenido para elaborar inventarios integrales. Sin embargo, aún constituye el método principal para elaborar pruebas de dominio específico, aunque casi siempre se complementa con otra información que apoya la validez de la prueba.

¡Inténtalo!

Emplea el método de contenido para elaborar reactivos para una prueba que mida "sociabilidad". En esta escala, una persona muy sociable tiene puntuaciones altas, y una no sociable tiene puntuaciones bajas. Se proporcionan los primeros dos reactivos; cada uno requiere una respuesta "Sí" o "No". Agrega otro reactivo.

	Sí	No
1. Disfruto estar con mucha gente.	O	O
2. Es difícil para mí conocer personas nuevas.	O	O
3. _____	O	O

Método de criterio meta

En el método de **criterio meta,** los reactivos de una escala de personalidad se seleccionan en términos de su capacidad para discriminar entre dos grupos bien definidos de examinados. En este contexto, "discriminar" se usa en el sentido de la discriminación del reactivo que revisamos en el capítulo 6 (véase pp. 148-150a»). Por lo común, uno de los grupos consta de individuos "normales", es decir, personas que se han identificado *sin* un padecimiento patológico, o como muestra representativa de la población general en la que la incidencia de cualquier padecimiento patológico es baja. El otro grupo, por lo general, está definido clínicamente por algún padecimiento patológico identificado con

claridad. Éste es el grupo "criterio" al que hace referencia el nombre de esta técnica, que también se conoce como método de meta empírica, ya que depende por completo de la definición empírica de cada escala.

Consideremos el siguiente ejemplo (basado en datos ficticios). Deseamos elaborar una prueba de personalidad que ayude a identificar a individuos deprimidos. Aplicamos los reactivos que aparecen en el cuadro 12-9 a un grupo de 50 personas que se han identificado con claridad como deprimidas por medio de una serie de entrevistas a profundidad con tres psicólogos. También aplicamos los reactivos a 50 individuos con un buen ajuste y no deprimidos. Los resultados de este estudio se resumen en el cuadro 12-9.

Cuadro 12-9. Datos ilustrativos (ficticios) del método de clave del criterio para elegir reactivos

Reactivo	(Marca Verdadero o Falso en cada reactivo)	%D	%N	Índice de Disc.	Elegido
1.	Estoy satisfecho con mi trabajo.	.67	.78	.11	
2.	Me siento incómodo con la gente.	.20	.18	.02	
3.	La mayoría de las personas es feliz.	.34	.67	.33	*
4.	Mi color favorito es el verde.	.31	.11	.20	*
5.	Me molestan muchas cosas.	.21	.18	.03	
6.	Me siento infeliz gran parte del tiempo.	.50	.10	.40	*
7.	Se puede confiar en la mayoría de las personas.	.78	.76	−.02	
8.	Me enfermo con frecuencia.	.32	.15	.17	*

%D = Porcentaje de examinados del grupo de deprimidos que marcaron Verdadero.
%N = Porcentaje de examinados del grupo normal que marcaron Verdadero.
Índice de Disc. = Índice de discriminación (%D − %N).
Elegidos (*) = Reactivos elegidos para ser incluidos en la escala de depresión.

En este ejemplo, elegiríamos los reactivos 3, 4, 6 y 8 para nuestra escala de depresión. La calificación de algunos reactivos está invertida, de modo que todas las respuestas "van en la misma dirección". Las puntuaciones bajas en nuestra escala indican depresión, aunque podríamos, obviamente, invertir la calificación de la escala entera de modo que las puntuaciones altas indicaran depresión. Hay varias sorpresas en los datos. Podríamos pensar que el reactivo 1 va en la escala de depresión; si usamos el método de contenido antes descrito, probablemente iría en dicha escala. Desde la perspectiva del contenido, el reactivo 5 también podría elegirse; sin embargo, en términos empíricos, ni el reactivo 1 ni el 5 discriminan de manera adecuada entre los grupos de depresión y normal, por lo que no son elegidos para nuestra escala. Los reactivos 3 y 6 se eligen para nuestra escala, lo cual no es sorpresivo; podemos entender por qué los individuos deprimidos responderían de manera diferente de los individuos normales en estos reactivos. Sin embargo, el reactivo 4 es una sorpresa. ¿Por qué el "color favorito" distingue entre personas deprimidas y normales? Aquí está la esencia del método de criterio meta: es irrelevante por qué este reactivo diferencia entre los grupos. El hecho es que lo hace; por lo tanto, es

un buen reactivo para la escala de depresión. Podemos comprender o no por qué el reactivo 8 diferencia, pero lo hace, por lo que también es seleccionado.

El método de clave del criterio se empleó por primera vez en la elaboración original del MMPI y el *Strong Interest Inventory*, dos pruebas que examinaremos con mayor detalle más adelante en este libro. En el trabajo del Strong, los grupos se definieron mediante la clasificación ocupacional (p. ej., trabajador social, plomero). Esta metodología funciona siempre que sea posible definir con claridad los grupos, y ahora está muy bien arraigada en la psicometría. De ahí que sea esencial que el estudiante de pruebas psicológicas la comprenda.

Ventajas y desventajas

El método de criterio meta ha demostrado ser sumamente provechoso en la elaboración de pruebas, pues centra su atención justo en lo que hace una prueba o en lo que queremos que haga. Su empirismo crudo es un antídoto útil para las nociones fantasiosas que, a menudo, se desbocan en el mundo de las teorías de la personalidad. Su franqueza y sencillez en la aplicación incentivan nuevas aplicaciones en la investigación. Éstas son las características positivas.

El método de criterio meta tiene tres principales inconvenientes. Primero, su orientación extremadamente ateórica limita la generalizabilidad de la interpretación de sus puntuaciones. Supongamos que demostramos que un conjunto de reactivos diferencia las personas deprimidas de las no deprimidas. ¿Qué más se puede hacer con la puntuación? Nada en el método de criterio meta sugiere alguna otra interpretación. ¿La puntuación se relaciona con algún rasgo general de personalidad? No. ¿La puntuación sugiere algo acerca de las características generales de la depresión? No, al menos no de manera evidente. ¿La puntuación proporciona alguna sugerencia para el tratamiento? No. Así, de no ser como auxiliar en la clasificación, el método de criterio meta es más bien estéril.

Segundo, el método de criterio meta se puede aplicar sólo cuando tenemos grupos criterio bien definidos. Tenemos tales grupos en el caso de categorías diagnósticas usadas de manera frecuente, como deprimido, paranoico, obsesivo, etc. Podemos crear tales grupos respecto de constructos como "ajuste" identificando por medio de nominaciones algunas personas que parezcan particularmente bien ajustadas y otras que no lo están. En este contexto, la clasificación puede no ser tan fácil como en el caso de los grupos clínicos mencionados. En el caso de otros rasgos de personalidad, puede ser bastante difícil identificar grupos criterio. Ejemplos de estos rasgos pueden incluir locus de control, autoestima y fortaleza del yo. Así, el método de criterio meta no parece ser igualmente aplicable a lo largo del espectro de los rasgos y características de personalidad que nos interesan.

El tercer punto acerca del método de criterio meta es, en realidad, una precaución relacionada con la interpretación más que una limitación del método por sí mismo. Éste hace hincapié en la diferenciación entre los grupos. La descripción típica del método, como la presentamos antes, hace hincapié en cómo esta diferenciación se maximiza.

Podríamos, con facilidad, tener la impresión de que el uso de este método conduce a una clara separación de los grupos, por ejemplo, entre individuos deprimidos y no deprimidos, que estaría definida con claridad por una puntuación de corte. Sin embargo, esto no ocurre así. La regla general es una superposición en la distribución de puntuaciones de los dos grupos más que una clara separación. En la interpretación de la prueba, necesitamos tener en mente el grado de superposición que existe en estas distribuciones. (En la discusión sobre grupos contrastados en el capítulo 5 se puede encontrar un tratamiento más detallado de este tema.)

Análisis factorial

Recordemos la explicación del **análisis factorial** que presentamos en el capítulo 5. El propósito básico es identificar las dimensiones (factores) que subyacen en una gran cantidad de observaciones, lo cual se logra examinando las interrelaciones (correlaciones) entre todas las observaciones. En el campo de las pruebas de capacidad, este proceso implica a menudo estudiar las correlaciones entre muchas pruebas, pero en el caso de las pruebas de personalidad, por lo general, implica examinar las correlaciones entre muchos *reactivos* de la prueba. Empezamos con un fondo grande de reactivos, como los que se muestran en el cuadro 12-2. Los reactivos se aplican a una muestra representativa de personas. El autor de la prueba determina las correlaciones entre las respuestas y luego recurre a la metodología analítico-factorial. Por último, el autor de la prueba interpreta los resultados, por lo general, proporcionando un nombre sucinto a cada factor y describiendo el significado de las dimensiones subyacentes.

Ventajas y desventajas

Es probable que existan algunas dimensiones básicas en la personalidad humana. El análisis factorial es la metodología primaria para ayudar a identificar estas dimensiones. Así, su principal ventaja es poner en orden una masa indiferenciada de reactivos y respuestas, pues al hacerlo se clarifica nuestro pensamiento acerca de la personalidad humana identificando "qué va con qué". Con base en la observación cotidiana, podemos debatir sin fin acerca de si las dos nociones representan las mismas características de personalidad o no. El análisis factorial ofrece un método empírico para responder estas preguntas, por lo que ha dado lugar a una gran cantidad de investigaciones. Es difícil que encontremos un número de una revista relacionada con la evaluación de la personalidad sin que aparezca el análisis factorial de alguna prueba de personalidad.

El método analítico-factorial en la elaboración de pruebas de personalidad tiene tres principales inconvenientes. Primero, los resultados finales dependen de manera crucial del contenido del fondo inicial de reactivos. Si no incluimos reactivos que se refieran a la dimensión X, el análisis factorial no identificará esa dimensión. Así, la lógica para generar el fondo inicial de reactivos es determinante.

Segundo, existe un debate interminable –y a menudo apasionado– entre los expertos en análisis factorial acerca de qué tan apropiadas son diferentes metodologías. ¿Qué tipo de coeficientes de correlación pueden o deben usarse? ¿Qué métodos de extracción de factores y qué procedimientos de rotación son apropiados? Tratar en detalle cualquiera de estas discusiones nos llevaría demasiado tiempo. Sin embargo, la frecuencia e intensidad de estos debates nos impulsa a ser cautelosos respecto de la aceptación superficial de cualquier aplicación única del análisis factorial.

Tercero, la descripción inicial del análisis factorial sugiere que podría producir un conjunto razonablemente definitivo de factores; sin embargo, la práctica real revela un cuadro diferente. Los resultados del trabajo analítico-factorial podrían describirse como fluidos más que definitivos. Por ejemplo, el *NEO Personality Inventory* (que se describe más adelante en este capítulo) se propone medir los cinco grandes rasgos de personalidad, pero, además, también produce puntuaciones de 30 "facetas", por lo que nos preguntamos: ¿hay cinco o 30 dimensiones básicas? El *Sixteen Personality Factor Inventory* (16 PF) identifica 16 dimensiones básicas de la personalidad humana, pero también produce cinco puntuaciones que corresponden a los cinco grandes factores. Otra vez, nos preguntamos: ¿hay 16 o cinco dimensiones básicas? Postular algún tipo de jerarquía entre los factores ofrece una manera de salir de esta dificultad. Este enfoque se ha desarrollado bien en el dominio de la capacidad, como vimos en el capítulo 7 con las teorías de la inteligencia, pero también es pertinente en el dominio de la personalidad.

Método teórico

El cuarto método de la elaboración de pruebas de personalidad depende de alguna teoría de la personalidad. El autor de la prueba adopta una teoría específica acerca de la personalidad humana y después construye los reactivos de la prueba, de modo que esta teoría se refleje en ellos. La teoría podría decir que existen cuatro rasgos dominantes o seis tipos de personas. Entonces, podría haber reactivos relacionados con cada uno de estos rasgos o tipos. La teoría podría ser ambiciosa e intentar describir la esfera total de la personalidad humana, o podría ser una teoría más restringida que intenta describir sólo uno de sus aspectos.

Ventajas y desventajas

La principal ventaja del método teórico es que ofrece una definición operacional de la teoría que debe conducir a más investigación acerca de ésta. La investigación, a su vez, puede llevar a un mayor desarrollo de la prueba. En las mejores circunstancias, la interacción de la elaboración de la prueba y la construcción teórica lleva a continuos refinamientos tanto en la prueba como en la teoría. Una buena teoría es un artefacto muy poderoso, y una buena prueba de una buena teoría puede ser de gran utilidad.

Existen dos inconvenientes principales del método teórico de la elaboración de pruebas.

Primero, la utilidad de la prueba, por lo general, está limitada por la validez de la teoría. Si ésta no es muy adecuada, entonces incluso una prueba que tiene una buena representación de la teoría no podrá ser de mucha utilidad. Segundo, siempre existe una preocupación acerca de qué tan bien la teoría se refleja en la prueba, incluso si la teoría es buena. No es fácil demostrar que una prueba específica es, en verdad, un reflejo válido de una teoría particular. Quizá por esta razón el enfoque teórico no se ha usado mucho en la elaboración de pruebas de personalidad como podría esperarse.

Métodos combinados

Como señalamos al principio de esta sección, comprender el método con que se elabora una prueba ayuda a comprender cómo puede usarse dicha prueba. Cada una sigue un método primario en su elaboración, el cual le da a la prueba cierto sello distintivo. Sin embargo, en la práctica, casi todas las pruebas emplean métodos múltiples en alguna etapa de su elaboración. Esto será más evidente cuando consideremos ejemplos específicos de pruebas más adelante en el capítulo, pero aquí señalamos algunos ejemplos de estos métodos múltiples.

Primero, al menos en cierto grado, todas las pruebas de personalidad empiezan con alguna versión del método de contenido. Para crear una lista inicial de reactivos, los autores de la prueba comienzan con interpretaciones comunes de los rasgos, características y patologías de la personalidad. Incluso en el método de criterio meta, en el que el contenido de los reactivos es inmaterial, los autores empiezan con reactivos que parecen relacionarse con las características de personalidad. Así, hasta cierto punto, todos usan el método de contenido. Segundo, en una veta relacionada, incluso cuando una prueba no está pensada como una medida directa de alguna teoría, la interpretación del creador de la prueba de las teorías de la personalidad sugerirá ciertos tipos de reactivos y escalas. Así, al menos hasta cierto punto, todos usan el método teórico.

Tercero, sin importar el método primario usado para elaborar la prueba, es casi inevitable que alguien lleve a cabo un análisis factorial de los reactivos. Los autores podrían hacerlo e informar los resultados en el manual de la prueba, pero también otros investigadores pueden hacerlo después de que la prueba se ha publicado e informar los resultados en un artículo publicado en una revista científica. Por último, lel criterio meta de los reactivos, para demostrar la diferenciación de dos o más grupos, se aplica, por lo común, a las pruebas de personalidad después de su publicación. Esto puede resultar en la sugerencia de escalas adicionales, sobre todo en el caso de las pruebas integrales de personalidad y su multiplicidad de reactivos. Así, tanto el análisis factorial como el criterio meta se aplican con frecuencia a las pruebas de personalidad incluso cuando estos métodos no son los medios primordiales con que se elaboran.

Ejemplos de inventarios integrales

En las siguientes secciones, examinaremos ejemplos de pruebas objetivas de personalidad; primero, algunos inventarios integrales y, después, algunas pruebas de dominio específico. Elegimos ejemplos que ilustran algunos puntos revisados en las secciones anteriores de este capítulo. No intentamos proporcionar una lista exhaustiva de todas las pruebas objetivas de personalidad, pues hay demasiadas para ello; además, hay algunas muy buenas fuentes de información con tales listas, como las que presentamos en el capítulo 2. En Camara, Nathan y Puente (1998, 2000), Hogan (2005a) y Piotrowski (1999) pueden consultarse listas con las pruebas más usadas. Asimismo, no intentamos hacer una revisión completa de las pruebas que presentamos, sino dar un panorama general de su naturaleza. En *Mental Measurements Yearbook* de Buros y en *Test Critiques* se pueden consultar dichas revisiones.

Edwards Personal Preference Schedule (EPPS): ejemplo de una prueba basada en la teoría

El *Edwards Personal Preference Schedule* [Inventario de Preferencias Personales de Edwards] (**EPPS**; Edwards, 1959) es un ejemplo de inventario objetivo de personalidad cuya estructura tiene su origen en una teoría de la personalidad, en este caso la de Henry Murray *et al*. (1938). El estudio real de 1938 de Murray, no obstante su amplia experiencia, ofreció una base notablemente frágil –un estudio con 50 (y sólo 50) hombres (y sólo hombres) en Harvard (y sólo en Harvard). Sin embargo, la publicación tuvo una influencia impresionante. Según parece, la teoría tocó una cuerda sensible de muchos psicólogos. El punto central es la noción de que, además de tener necesidades biológicas básicas (Murray identificó 12 de estas necesidades "viscerogénicas"), la gente tiene necesidades psicológicas básicas. Es difícil discernir una lista única y definitiva de estas necesidades psicológicas en su trabajo original, pero había al menos 20. Por ejemplo, Engler (1999) propuso 20 necesidades, y Carver y Scheier (1996), 27. Algunas de las necesidades más reconocidas son las de Logro (necesidad de ser exitoso) y Afiliación (necesidad de tener amistades y otras relaciones agradables). Edwards se propuso medir 15 de estas necesidades; en el cuadro 12-10 aparecen algunos ejemplos. Por lo general, cuando una prueba toma su estructura de una teoría particular, el manual de la prueba hace un esfuerzo para explicar la teoría y demostrar de qué manera se refleja en la prueba. Esto no ocurre con el EPPS; después de un breve reconocimiento de que la prueba pretende medir variables de la lista de Murray, no hay otras menciones de la teoría ni una explicación de por qué sólo se eligieron 15 necesidades, ni algún intento para demostrar que los reactivos representan de manera equitativa cada necesidad.

Cuadro 12-10. Ejemplos de escalas en el Edwards Personal Preference Schedule

Escala	Breve descripción
1. Logro	Necesidad de esforzarse y dar lo mejor de sí
2. Deferencia	Necesidad de recibir sugerencias de otros y amoldarse a ellos
3. Orden	Necesidad de ser ordenado y organizado
4. Exhibición	Necesidad de llamar la atención y ser reconocido

El manual del EPPS expresa una gran preocupación acerca de la influencia de la deseabilidad social en las respuestas. Para tratar con esta influencia, la prueba emplea una metodología de elección forzada en los reactivos, por lo que el EPPS ofrece un excelente ejemplo de las consecuencias de usar este formato de reactivos de elección forzada. Regresaremos a este tema.

El EPPS consta de 225 *pares* de afirmaciones, la mayoría de las cuales empieza con "Siento..." o "Me gusta..." Cada afirmación representa una de las 15 necesidades; cada una se aparea con todas las otras necesidades, lo que resulta en 14 pares para cada una de las 15 necesidades; entonces, hay dos conjuntos de estos pares de las 15 necesidades. Así, un examinado que *siempre* escoge las afirmaciones que representan, digamos, Logro recibe una puntuación natural de 28 en Logro. Un examinado que nunca escoge las afirmaciones de Afiliación obtiene una puntuación natural de 0 en Afiliación. El cuadro 12-11 muestra un ejemplo de un reactivo del EPPS. La afirmación A representa la necesidad de Afiliación, mientras que la afirmación B representa la necesidad de Logro.

Cuadro 12-11. Reactivo muestra del EPPS
76 A. Me gusta ser leal con mis amigos. B. Me gusta hacer mi mejor esfuerzo en cualquier cosa que emprendo.
Fuente: *Edwards Personal Preference Schedule* (Edwards, 1959). Con autorización de Allen L. Edwards Living Trust.

La metodología de elección forzada produce **puntuaciones ipsativas**. Si se obtiene una puntuación alta en algún área (necesidad), necesariamente se obtiene una puntuación baja en otra; no es posible tener puntuaciones altas o bajas en todas las escalas. La puntuación *promedio* a lo largo de todas las escalas siempre es la misma para todas las personas. Si tu puntuación más alta es en Afiliación, significa que es mayor que en cualquier otra área, pero no que eres más alto en Afiliación en un sentido absoluto ni más alto que otras personas en Afiliación, aunque a veces puede ser así. Si tu puntuación más baja es en Logro, no necesariamente significa que seas más bajo en Logro en un sentido absoluto, sino sólo que eres más bajo en Logro que en otras áreas. La interpretación de las puntuaciones ipsativas requiere cierta gimnasia mental, sobre todo cuando las puntuaciones naturales ipsativas se traducen en normas, como ocurre en el EPPS.

Además de proporcionar puntuaciones de 15 necesidades, el EPPS produce puntuaciones de Consistencia y Estabilidad del perfil. La puntuación de Consistencia se basa en 15 pares de reactivos que están repetidos de manera exacta en la prueba. El individuo típico (medio) es consistente en al menos 12 pares. El manual del EPPS

sugiere que una puntuación de Consistencia de 9 (sólo 2% de la muestra universitaria de estandarización tuvo puntuaciones menores de 9) lleva a cuestionar la validez de la prueba entera. Éste es un buen ejemplo de un índice de validez de una prueba objetiva de personalidad.

La puntuación de Estabilidad del perfil se determina, de manera ingeniosa, por la manera en que las respuestas se registran en la hoja de respuestas. La mitad de las respuestas de cada variable se ubica en una hilera y la otra mitad, en una columna. La Estabilidad del perfil se determina correlacionando las puntuaciones de las hileras con las de las columnas.

El EPPS es el ejemplo perfecto de una prueba con un comienzo promisorio, pero con un seguimiento y una revisión inadecuados. Quizá por esta razón el EPPS ha caído en desuso. Sin embargo, merece estar incluido aquí, porque ilustra muy bien el origen de una prueba en una teoría, así como el uso del método ipsativo de medición.

NEO Personality Inventory-3: ejemplo de una prueba analítico-factorial «323-326a

Los psicólogos tienen una larga tradición aplicando el análisis factorial al campo de la personalidad. Entre los pioneros de esta tradición están Guilford (1959a), Cattell (1966) y Eysenck (1970); todos ellos han creado pruebas basadas en el análisis factorial de la personalidad. Además, incluso las pruebas que se elaboraron desde otra perspectiva, por ejemplo, el MMPI o el EPPS, han sido objeto, con frecuencia, del análisis factorial. De todas estas investigaciones surgió la teoría de los "Cinco Grandes" o de los "cinco factores" de la personalidad. En un artículo muy citado, Digman (1990, p. 418) se refiere a la teoría de los Cinco Grandes como "una rápida convergencia de puntos de vista relacionados con la estructura de los conceptos de la personalidad... [y] ... una estructura teórica de sorprendente generalidad". Wiggins y Trapnell (1997) se refieren al modelo de los cinco factores como "venerable" y representante de "consenso de trabajo" en el campo de la teoría de la personalidad. Ozer y Benet-Martinez (2006) señalaron que "las amplias dimensiones supraordinadas... de los modelos de los cinco factores de la personalidad... se usan mucho ahora en la literatura de la personalidad y la predicción" (p. 402).

De manera muy sencilla, la teoría dice que, considerando toda la investigación analítico-factorial, parece que existen cinco factores o dimensiones básicas en la personalidad humana. Ésta es una "teoría" sólo en el sentido más débil, pues una teoría debe tener poder explicativo, mostrar las relaciones dinámicas entre los constructos y formular predicciones.

La teoría de los cinco factores no hace nada de esto: es puramente descriptiva. Dice que la investigación, por lo general, identifica cinco factores de personalidad; de ahí que a veces se denomine modelo más que teoría. No obstante, en el largo, arduo y controvertido estudio de la personalidad, esta generalización acerca de los Cinco Grandes representa un hito importante.

Los nombres exactos de los cinco factores varían un poco de un autor a otro. Un sistema que se suele usar aplica estos nombres: Franqueza, Escrupulosidad, Extroversión, Simpatía y Neuroticismo.

Debemos señalar que la investigación más reciente sugiere que puede haber un sexto factor en la personalidad humana. Se trata del sentido ético básico, que incluye honestidad, humildad e integridad, y a veces se le denomina Honestidad/Decencia (Saucier, 2009; Thalmayer, Saucier, & Eigenhuis, 2011). Así, ahora tenemos un modelo de los Seis Grandes que compite con el de los Cinco Grandes, si bien este último aún predomina en la literatura profesional.

El *NEO Personality Inventory-3* [Inventario de Personalidad NEO] (**NEO PI**: McCrae & Costa, 2010) es considerado una de las medidas principales de los Cinco Grandes factores de personalidad; de ahí que lo usemos como nuestro ejemplo principal del método analítico factorial para elaborar pruebas de personalidad. Los "inventarios NEO" constituyen una familia de medidas relacionadas e informes. Las versiones más recientes incluyen el NEO PI-3 en dos versiones, cada una con 240 reactivos: la Forma S de autoinforme (es decir, el examinado se valora a sí mismo) y la Forma R, en la que se informa sobre otra persona (con cuadernillos separados para hombres y mujeres), el NEO FFI-3, una versión corta de 60 reactivos del NEO PI-3, y varias formas de informe. Aquí describimos la Forma S del NEO PI-3, la más usada de todas las versiones.

El NEO PI-3 deriva de su predecesor, el NEO PI-R, con modificaciones en 37 de los 240 reactivos, principalmente con el fin de simplificar la carga de lectura, lo cual hace aplicable la prueba a grupos más jóvenes. La nueva versión mantiene la misma estructura exacta de la versión anterior; tiene puntuaciones de cinco dominios que corresponden a los cinco constructos del modelo de los Cinco Grandes. Además, cada dominio tiene seis escalas de facetas que representan características más finas, de modo que se obtienen 30 puntuaciones de facetas. El cuadro 12-12 bosqueja esta estructura.

Cuadro 12-12. Lista de dominios y facetas del NEO PI

Dominios	**Facetas de Extroversión (E)**	**Facetas de Simpatía (S)**
N: Neuroticismo	E1: Calidez	S:1 Confianza
E: Extroversión	E2: Sociabilidad	S2: Claridad
F: Franqueza	E3: Asertividad	S3: Altruismo
S: Simpatía	E4: Actividad	S4: Sumisión
Es: Escrupulosidad	E5: Búsqueda de emociones	S5: Modestia
	E6: Emociones positivas	S6: Bondad
Facetas de Neuroticismo (N)	**Facetas de Franqueza (F)**	**Facetas de Escrupulosidad (Es)**
N1: Ansiedad	F1: Fantasía	Es1: Competencia
N2: Hostilidad Enojo	F2: Estética	Es2: Orden
N3: Depresión	F3: Sentimientos	Es3: Obediencia
N4: Timidez	F4: Acciones	Es4: Búsqueda de logros

N5: Impulsividad	F5: Ideas	Es5: Autodisciplina
N6: Vulnerabilidad	F6: Valores	Es6: Deliberación

Reproducido con el permiso especial de la editorial Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, de *NEO Inventories Profesional Manual*, por Paul T. Costa, Jr.,PhD. y Robert R. McCrae, PhD. Copyright © 2010 por Psychological Assessment Resources, Inc.,, Inc. Prohibida cualquier otra reproducción sin la autorización de PAR, Inc.

Los troncos de los reactivos en el NEO PI son similares a los reactivos genéricos del cuadro 12-2. El examinado responde en una escala de cinco puntos: Totalmente de acuerdo, De acuerdo, Neutral, En desacuerdo y Totalmente en desacuerdo. La prueba produce *cinco* puntuaciones de *dominio*, que corresponden a las Cinco Grandes características de personalidad. Cada dominio contiene *seis* puntuaciones de *facetas*, las cuales representan manifestaciones más específicas de los dominios. Con cinco dominios y seis facetas por dominio, la prueba tiene un total de 30 puntuaciones de facetas. Cada faceta se mide con ocho reactivos, lo cual significa que hay 240 reactivos en total (8 × 30). El cuadro 12-12 muestra los nombres de los dominios y facetas del NEO PI; por cada faceta, el manual de la prueba proporciona una breve descripción, por lo general con un contraste entre las puntuaciones altas y bajas. El cuadro 12-13 presenta la descripción del manual de la faceta Asertividad, que pertenece al dominio de Extroversión.

Cuadro 12-13. Descripción de la faceta de Asertividad del dominio Extroversión del NEO PI

E3: *Asertividad*. Quienes obtienen puntuaciones altas en esta faceta son dominantes, enérgicos y presentan ascendencia social. Hablan sin titubear y, a menudo, se convierten en líderes grupales. Quienes obtienen puntuaciones bajas prefieren mantenerse en segundo plano y dejar que otros dirijan la conversación.

Fuente: Reproducido con el permiso especial de la editorial PAR, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, de *NEO Inventories*, por Robert R. McCrae, PhD. y Paul T. Costa, Jr., PhD. Copyright © 2010 por PAR, Inc. Prohibida cualquier reproducción sin la autorización de PAR, Inc.

El NEO PI no tiene ningún índice de validez formal. Se pregunta a los examinados, al final del inventario, si han respondido todas las preguntas y si lo hicieron con honestidad. Evidentemente, si un examinado indica que no respondió con honestidad, la validez de la prueba es cuestionable. El manual también recomienda la inspección visual de las respuestas para determinar un patrón (p. ej., diez "Totalmente de acuerdo" consecutivas) que pueda indicar respuestas no sinceras. El manual recomienda contar el número de respuestas De acuerdo y Totalmente de acuerdo para detectar tendencias a decir sí a todo (cuenta alta) o no a todo (cuenta baja). Sin embargo, no se ofrecen normas para este conteo.

Normas. El NEO PI-3 proporciona normas; las puntuaciones principales son convertidas en puntuaciones T para adolescentes (de 14 a 20 años de edad) y adultos (de más de 21 años). El manual presenta ambos conjuntos de normas por separado para

hombres y mujeres y para los géneros combinados. Las normas de adolescentes se basan en 500 casos, y las de adultos en 635 casos. El manual ofrece escasa información acerca de estos grupos de estandarización. Siguiendo las referencias del manual a informes tempranos sobre estos grupos (McCrae, Martin, & Costa, 2005), se encuentra con claridad que ninguno de los dos fueron representativos de la población de adolescentes y adultos de EUA. Por ejemplo, la muestra de adultos provino en su mayoría (63%) de sólo un estado, 93% fue blanca y más de la cuarta parte tenía un grado superior al de licenciatura. La muestra de adolescentes tuvo 2% de hispanos y 3% de negros (en comparación con aproximadamente 13% de cada grupo en la población nacional), y dos terceras partes de la muestra esperaban alcanzar un grado superior al de licenciatura. El manual también presenta diversos conjuntos de normas "complementarias".

Confiabilidad. Basados en la muestra de adultos descrita, los coeficientes de consistencia interna (alpha) de los cinco dominios caen en un estrecho rango que va de .89 a .93, con una mediana de .90. Los coeficientes alpha de las 30 facetas varían entre .54 y .83, con una mediana de .76. Así, las puntuaciones de los dominios demuestran una buena confiabilidad de consistencia interna, mientras que las confiabilidades de las puntuaciones de facetas son mediocres. ¿Tenemos que decirlo otra vez? Con muchos reactivos, por lo general, se obtiene una buena confiabilidad; con pocos reactivos, tendremos problemas. Los datos de la confiabilidad de test-retest provienen de un estudio relativamente pequeño (N = 132) con el NEO PI-R, no con el NEO PI-3. Los resultados son parecidos a los de los coeficientes alpha: buena confiabilidad en los dominios y marginal en muchas facetas.

Validez. El número de estudios de validez del NEO PI, si incluimos los que se basan en todas sus ediciones y variaciones, es impresionante: asciende a cientos y, quizá, a miles. El NEO PI-3 difiere poco en estructura o reactivos con respecto a su predecesor, el NEO PI-R, así que es razonable suponer que la validez de una prueba es generalizable a la otra. La mayoría de estudios encaja en tres categorías amplias. Primero, hay estudios que confirman la estructura factorial del inventario. Segundo, hay estudios que correlacionan el NEO PI con otras escalas de personalidad. Tercero, hay estudios de contrastes grupales, es decir, comparaciones de las puntuaciones en el NEO PI de un grupo frente a otro, o de un grupo frente al grupo de estandarización. En general, los estudios apoyan la validez del NEO PI como una medida de los cinco principales constructos meta. En Botwin (1995), Juni (1995) y Tinsley (1994) se pueden encontrar reseñas del NEO PI-R.

Cuando se califica de manera automatizada, el NEO PI produce un informe interpretativo extenso. El cuadro 12-14 ilustra cómo puede crearse una breve porción del informe sobre una persona ficticia. Las afirmaciones en el informe dependen, en parte, de la simple traducción de las comparaciones normativas y, en parte, de las implicaciones de las investigaciones con la prueba. En el capítulo 3, p. 65a», se puede encontrar un ejemplo de informe interpretativo.

Cuadro 12-14. Ejemplos de afirmaciones que podrían aparecer en el informe interpretativo del NEO PI

Lo que aparece en el informe interpretativo	Comentarios sobre la base de la afirmación
Tus puntuaciones en los dominios del NEO Pi son N E F S Es 48 59 52 55 45	Éstas son puntuaciones T. Aparecen en forma numérica, así como en gráficas.
El NEO PI es una medida de rasgos comunes de personalidad dentro del rango normal de variabilidad. El informe muestra tu estatus en los cinco rasgos principales.	Esto viene en una "plantilla"; todos lo obtienen.
La característica más distintiva de tu perfil de puntuaciones es la puntuación relativamente alta en Extroversión.	Esta afirmación surge directamente del perfil de puntuaciones T.
Las personas con puntuaciones en Extroversión como la tuya tienden a ser muy sociables. Les gusta estar con otras personas e interactuar con ellas. Prefieren eso a trabajar o estar solas.	Esta afirmación proviene de la investigación sobre las características de quienes obtienen puntuaciones altas en "E", aunque, hasta cierto punto, surge simplemente de la definición factorial de la escala de Extroversión.
Etc. - casi 15 páginas más (!) que incluyen las *Gráficas de Estilo*	Continúa con otras afirmaciones basadas en normas e investigación, a menudo relacionadas más con "plantillas".

¡Inténtalo!

Puedes encontrar sin dificultad copias enteras de informes interpretativos del NEO PI en internet. Escribe "NEO PI interpretive report" [informe interpretativo del NEO PI] en cualquier buscador. Consulta uno de los informes. ¿Puedes identificar qué partes del informe se basan en comparaciones normativas y cuáles en plantillas predeterminadas?

Todas las versiones de los inventarios NEO están dirigidas al rango normal de los rasgos de personalidad más que a los padecimientos patológicos (p. ej., depresión, pensamientos distorsionados). No obstante, el NEO PI-3 coquetea con el lado anormal de las cosas con un nuevo *Problems in Living Checklist* [Lista de chequeo de problemas en la vida] (PLC) que está claramente dirigido a usuarios clínicos. El PLC no es una prueba separada, sino que usa puntuaciones del NEO PI, sobre todo las puntuaciones extremas, para sugerir vías para el seguimiento clínico. El informe interpretativo del NEO PI también contiene una sección sobre los trastornos del Eje II del DSM-IV.

Los informes del NEO PI también ofrecen "Gráficas de Estilo", es decir, grafican en un sistema bidimensional de coordenadas las puntuaciones de cada *par* de dominios (así, 10 gráficas en total, p. ej., N y E, N y F). Las categorías descriptivas adornan segmentos del espacio coordinado, que recuerda de manera vaga los muy citados indicadores de tipo Myers-Briggs.

IPIP: Construye tu propio inventario de personalidad

Como regla general, presentamos en este libro sólo las pruebas que se publican con regularidad, pero hacemos una excepción con esta entrada: el *International Personality Item Pool* [Fondo internacional de reactivos de personalidad], que se conoce popularmente como IPIP y puede encontrarse en la dirección http://ipip.ori.org. El IPIP no es una prueba específica, sino, como lo sugiere su título, un fondo de reactivos del dominio de la personalidad con una orientación hacia los Cinco Grandes rasgos de la personalidad, como los que encontramos en el NEO PI. El IPIP contiene más de 2000 reactivos relacionados con la personalidad, varias escalas compuestas de reactivos del fondo, instrucciones para aplicar y calificar, y referencias cruzadas de los reactivos y los Cinco Grandes rasgos de personalidad. Los reactivos y las escalas están disponibles gratuitamente para su uso sin que sea necesario pedir permiso; por ello se han usado mucho en las investigaciones. Los reactivos y las escalas del IPIP no se usan mucho en la práctica clínica profesional.

Pruebas de dominio específico

Ahora examinaremos medidas de dominio específico, pero antes, quizá, el lector desee revisar las características generales de estas medidas como se resumen en el cuadro 12-6. Literalmente, existen miles de estas medidas. Examinaremos una prueba específica y, después, hablaremos del área que está surgiendo donde las pruebas de dominio específico se usan con frecuencia.

Piers-Harris Children's Self-Concept Scale

El autoconcepto es un constructo de mucho interés para los psicólogos, por lo que se han creado numerosas pruebas para medirlo. Aquí describimos el *Piers-Harris Children's Self-Concept Scale* [Escala de Autoconcepto Infantil de Piers-Harris], Segunda Edición (Piers & Herzberg, 2002). Con el subtítulo "The way I feel about myself" [Cómo me siento conmigo mismo], la prueba, por lo general, se conoce simplemente como el **Piers-Harris** 2, aunque el acrónimo PHCSCS, difícil de manejar, también se encuentra a veces. De las numerosas medidas existentes de autoconcepto, quizá ésta es la que más se usa; al menos en su primera edición (Piers, 1996), estuvo clasificada entre las medidas más usadas en campos como la psicología escolar (Archer, *et al*., 1991; Kennedy *et al*., 1994; Piotrowski, 1999).

El Piers-Harris 2 es una medida de autoinforme de 60 reactivos dirigida a niños y adolescentes de 7 a 18 años de edad. El examinado responde a afirmaciones acerca de sí mismo en un formato de Sí o No. En el cuadro 12-15 se pueden ver dos afirmaciones muestra. Casi la mitad de los reactivos está redactada de modo que la respuesta Sí indica un autoconcepto positivo (p. ej., el reactivo 18 del cuadro 12-15); la otra mitad está redactada de modo que No indica un autoconcepto positivo (p. ej., el reactivo 29). Este arreglo ofrece un ejemplo del equilibrio en la direccionalidad de los reactivos para controlar las tendencias a decir sí o no a todo.

Cuadro 12-15. Reactivos parecidos a los del Piers-Harris

	Sí	**No**
Tengo buenas calificaciones en la escuela.	S	N
Me altero con facilidad.	S	N

El Piers-Harris 2 produce una puntuación total y seis puntuaciones de "dominio" derivadas del análisis factorial de los reactivos. El cuadro 12-16 muestra los nombres de los dominios y el número de reactivos por dominio. Muchos reactivos aparecen en más de un dominio; algunos se encuentran hasta en tres de ellos. Esto es muy extraño dado el origen analítico-factorial de los dominios.

Cuadro 12-16. Puntuaciones del Piers-Harris

Dominio	**No. de reactivos**
Adaptación conductual (CON)	14
Estatus escolar e intelectual (INT)	16
Apariencia y atributos físicos (FIS)	11
Libre de ansiedad (LIB)	14
Popularidad (POP)	12
Felicidad y satisfacción (FEL)	10
Total	60[a]
Índice de inconsistencia	15
Índice de Sesgo de la respuesta	(todos los reactivos)

[a] El número total de reactivos es menor que la suma del número de reactivos dentro de los dominios, porque algunos aparecen en más de uno.

Un índice de inconsistencia se deriva comparando las respuestas de 15 pares de reactivos relacionados en términos conceptuales y empíricos. El Índice de Sesgo de la respuesta se basa en una simple suma de todas las respuestas "Sí", de modo que varía de 0 a 60. Una puntuación baja o alta puede indicar una tendencia a decir no o sí a todo, respectivamente. El manual del Piers-Harris sugiere puntuaciones de corte para identificar patrones inconsistentes de respuesta y tendencias a decir sí o no a todo; estos índices también se grafican como variables continuas en los perfiles de puntuaciones. Éstos son excelentes ejemplos de los índices de validez de una prueba de dominio específico. En el caso de la deseabilidad social, no hay un índice de validez; el manual sólo señala la importancia de tomar en cuenta esta variable al interpretar las puntuaciones.

Las puntuaciones naturales del Piers-Harris –los dos índices de validez, los seis dominios y la puntuación total– se grafican en un perfil que muestra las puntuaciones T (normalizadas) y los percentiles. Las normas se basan en una muestra de 1387 niños con una razonable representatividad de la población de EUA por sexo, antecedentes étnicos/raciales, región geográfica y nivel de educación de la familia nuclear. La muestra normativa del Piers-Harris 2 es una mejora importante en comparación con la muestra de la primera edición, pues en aquella ocasión la muestra se limitó a sólo un sistema escolar rural.

El perfil de puntuaciones se puede graficar a mano. La editorial también produce tres informes interpretativos cuando la prueba se califica de manera automatizada. El más detallado de éstos, el Informe individual, produce comentarios interpretativos de cada puntuación que se presenta en el cuadro 12-16. Tomando en cuenta que se trata de una prueba de dominio específico, los informes interpretativos del Piers-Harris 2 son extraordinariamente extensos.

La confiabilidad de consistencia interna (alpha) del Piers-Harris 2, basada en la muestra de estandarización, tiene un promedio de casi .90 en la puntuación total y .78 en las puntuaciones de dominio; sólo algunos índices caen por debajo de .75, incluso de .69, en

ciertos grupos de edad. En el Piers-Harris 2 no se presentan datos de la confiabilidad de test-retest; en el caso de la primera edición, estos datos se limitaron casi exclusivamente a las puntuaciones totales. Ésta es una deficiencia técnica importante del manual.

La información de validez que se presenta en el manual consta primordialmente de correlaciones con las valoraciones de los maestros y compañeros, correlaciones con una gran cantidad de otras pruebas, contrastes entre grupos que se piensa tienen un autoconcepto alto o bajo y la consideración de la estructura factorial de la prueba. Los estudios analítico-factoriales evocan preguntas interesantes. La intercorrelación mediana entre los dominios es de .64, aunque algunas correlaciones alcanzan más de .80. Éstas son correlaciones altas para dominios que, se supone, son relativamente independientes. Hasta cierto punto, estas intercorrelaciones se deben al hecho de que algunos reactivos aparecen en más de una escala. La revisión de Oswald (2005) del Piers-Harris 2 planteó tales dudas acerca de la estructura factorial de la prueba; sin embargo, de manera más amplia, debemos encarar la cuestión de si el autoconcepto tiene una estructura jerárquica, como la inteligencia (véase capítulo 7). Marsh ha argüido, de manera enérgica, que las correlaciones entre diferentes facetas del autoconcepto son tan bajas que no es legítimo producir una puntuación "total" (véase Marsh, Parada, & Ayotte, 2004).

Aquí está el meollo del argumento: si varias subpruebas tienen correlaciones muy altas entre sí, entonces parece que sólo existe un rasgo subyacente (en este caso, el autoconcepto general), por lo que las subpuntuaciones separadas no deberían obtenerse. Por otro lado, si varias subpuntuaciones tienen correlaciones muy bajas, no tiene sentido sumarlas para obtener una puntuación total, por lo que deberían permanecer separadas. Por ejemplo, ¿qué sentido tiene proporcionar una "puntuación total" de estatura, CI y velocidad en la carrera de 100 metros (tres rasgos que tienen una correlación cercana a 0)? Este argumento ofrece un buen ejemplo de cómo una teoría sólida, basada en la investigación psicométrica, puede y debe informar el uso práctico de las pruebas.

El Piers-Harris 2, la edición que aquí describimos, difiere ligeramente de la primera en la estructura. La segunda edición tiene menos reactivos (60 frente a 80), pero cubre las mismas áreas básicas, que en este contexto se denominan dominios y no grupos, y los nombre de los dominios se ajustaron ligeramente. La diferencia más grande entre ambas ediciones es la base normativa, como señalamos antes. En Kelly (2005) y Oswald (2005) se pueden encontrar revisiones del Piers-Harris 2.

Medidas en el campo de la psicología positiva

Este capítulo se concentra en medidas dentro del rango normal de los rasgos de personalidad y deja aparte las medidas de padecimientos clínicos, que serán tratadas en el siguiente capítulo. El contraste es quizá mayor cuando consideramos medidas dentro de la psicología positiva.

Las medidas consideradas en el siguiente capítulo se centran en la conducta problemática y los estados cognitivos: depresión, ansiedad, obsesión, etc. Esto es comprensible porque son la materia del trabajo clínico. Sin embargo, el reciente

desarrollo de la **psicología positiva** reclama nuestra atención para el otro lado de los seres humanos, como si sugiriera que la psicología del siglo XXI pudiera tratar de manera más satisfactoria este otro lado: en el número de enero de 2000 de *American Psychologist*, la revista emblemática de la *American Psychological Association*, estuvo dedicado, de manera especial, a la felicidad, la excelencia y el funcionamiento humano óptimo. Como señalaron Seligman y Csikszmenthalyi (2000) en un artículo pionero: "Enfocarse exclusivamente en la patología, lo cual ha predominado tanto en nuestra disciplina, resulta en un modelo del ser humano carente de características positivas que hacen que la vida valga la pena" (p. 5). En este campo en crecimiento se encuentran constructos como esperanza, valor, prudencia, vitalidad, sabiduría, optimismo, humor, espiritualidad y muchos otros, además de la dominante noción de felicidad. Peterson y Seligman (2004) proporcionaron una taxonomía preliminar de constructos pertenecientes a este campo y subtitularon la introducción como "manual de la sensatez" (p. 13). La taxonomía tiene seis categorías principales y 24 subcategorías.

Los campos que recién surgen a menudo dejan la medición para más tarde, pues prefieren concentrar su energía en la descripción inicial de constructos pertinentes y sus posibles interrelaciones. Es mérito de quienes están activos en el campo de la psicología positiva no haber demorado el trabajo de medición, pues su trabajo está lleno de intentos de proporcionar definiciones operacionales de sus constructos por medio de pruebas específicas. Por ejemplo, cada capítulo del volumen editado por Lopez y Snyder (2003) considera varias medidas de una gran cantidad de constructos de la psicología positiva, por ejemplo, gratitud, empatía y sentido del humor. Además, cada capítulo de la taxonomía de Peterson y Seligman (2004) se concentra en la descripción de un constructo, pero también se refiere a las medidas de cada uno de ellos.

Los lectores interesados en una exploración más amplia de las medidas de constructos de la psicología positiva deben consultar los libros de Lopez y Snyder (2003) y Peterson y Seligman (2004). Por ahora, nos limitamos a hacer tres generalizaciones de estas medidas. Primero, como en el caso de las medidas populares que examinamos (p. ej., NEO-PI, MMPI-2), la mayoría de las medidas de los constructos de la psicología positiva es de la variedad de autoinforme, en el que se usan afirmaciones sencillas como reactivos y formatos de respuesta tipo Likert. Regresa a los formatos de respuesta del cuadro 12-1 y a las afirmaciones del cuadro 12-2, usa los mismos tipos de formatos de respuesta y sustituye con las afirmaciones apropiadas para el constructo pertinente: esto es lo que encuentras en una medida típica de constructos de la psicología positiva. Desde luego, existen algunas excepciones.

La segunda generalización se relaciona con los tipos de investigación psicométrica que se llevan a cabo con estas medidas. Las cuestiones son las mismas que con otras medidas: consistencia interna y confiabilidad de test-retest, validación mediante el análisis factorial, examen de las diferencias grupales esperadas, estudios de validez discriminante y convergente, y, desde luego, la siempre presente preocupación por la representatividad de los datos normativos.

Tercero, mientras que algunas medidas de personalidad intentan cubrir muchas áreas

diferentes, por ejemplo, la multiplicidad de puntuaciones que se obtienen del NEO PI o el MMPI-2, las medidas en la psicología positiva tienden a concentrarse sólo en un constructo o en algunos que están estrechamente relacionados. No existen inventarios integrales en la psicología positiva, al menos no aún. La taxonomía de Peterson-Seligman puede ofrecer una base para algún cambio en esta circunstancia.

Cuadro 12-17. Ejemplos de rasgos de la psicología positiva

Optimismo	Esperanza	Autoeficacia
Gratitud	Perdón	Empatía
Autoestima	Sabiduría	Bienestar subjetivo
Valor	Prudencia	Humildad

¡Inténtalo!

Considera los rasgos identificados en el cuadro 12-17. Éstos son sólo ejemplos, no una lista exhaustiva. ¿Puedes agregar alguno?

Ejemplo: bienestar subjetivo

¿En general, qué tan satisfecho estás con tu vida? Responde en una escala de 1 a 10, en la que 1 es muy insatisfecho y 10 es muy satisfecho.

¿Qué tan feliz estás? Responde en una escala de 1 (muy infeliz) a 5 (muy feliz).

Marca la cara que muestra como te sientes con mayor frecuencia.

Estos reactivos ilustran los métodos para medir el concepto global de bienestar subjetivo (BS). Aunque la psicología positiva es relativamente nueva como campo identificable, el BS se ha estudiado por muchos años, pero se ajusta muy bien a este nuevo campo. La medición del BS presenta algunos retos y resultados muy interesantes.

En términos de estructura general, el BS sigue un patrón muy similar al del modelo jerárquico de la capacidad mental examinado en el capítulo 7. En el nivel más bajo, hay muchos sentimientos específicos y evaluaciones personales; por ejemplo, ¿cómo te sientes en relación con esta asignatura escolar en comparación con aquélla? ¿Cómo te sientes el viernes en la tarde en comparación con el lunes en la mañana? Estas evaluaciones sumamente específicas se conjuntan en varios niveles intermedios. Keyes y Magyar-Moe (2003) bosquejaron varios sistemas existentes. Uno distingue entre el bienestar psicológico y el bienestar social; estas dos categorías tienen, a su vez, subdivisiones. Por ejemplo, el bienestar psicológico comprende áreas como

autoaceptación, propósito en la vida y crecimiento personal. Otro sistema distingue entre bienestar psicológico, social y emocional, cada uno con varias subcategorías. Power (2003) resumió el sistema empleado en el estudio Calidad de vida de la Organización Mundial de la Salud. Este sistema tenía cuatro dominios (físico, psicológico, social y ambiental), cada uno con diversas facetas. Existen instrumentos de medición multidimensional para cada uno de estos sistemas.

Igual que el modelo jerárquico de la capacidad mental tenía "g" en la cima, los sistemas que acabamos de describir parecen equivaler a un BS global, que se evalúa con escalas de afirmaciones y respuestas muy sencillas que presentamos al principio de esta sección. Se ha realizado una gran cantidad de investigación (muchos estudios transculturales) que emplea estas medidas sencillas. En Diener (2000) se puede encontrar resúmenes del uso de estas medidas globales y sus correlaciones con una gran variedad de conductas del mundo real. En Diener (1984), Eid y Larsen (2008), Myers (2000), Peterson (2006), Ryan y Deci (2001) y Seligman *et al*. (2005) se puede encontrar informes relacionados.

Una ramificación fascinante de la investigación sobre el BS es la clasificación de países enteros en términos de la "felicidad promedio" de los ciudadanos. Al igual que en la clasificación de países en materias como aprovechamiento en ciencia y matemáticas (véase capítulo 11, p. 301a»), también tenemos clasificaciones de países en medidas variadas del BS. Una fuente agradable es Helliwell, Layard y Sachs (2012). Escribe "felicidad mundial" en un buscador de internet para encontrar varias de estas clasificaciones.

Resumen de puntos clave 12-4

Tendencias de las pruebas objetivas de personalidad
Se publican muchas pruebas nuevas
Han madurado los métodos de elaboración
La administración cuidadosa enfatiza la conveniencia y los bajos costos
Ahora es común la producción de informes interpretativos
Se desarrolla con rapidez la aplicación en línea

Tendencias en la elaboración y uso de las pruebas objetivas de personalidad

¿Qué conclusiones podemos formular de nuestra inspección de las pruebas objetivas de personalidad? ¿Y qué podemos prever en su futuro? Aquí presentamos los siguiente puntos en respuesta a estas preguntas (véase Resumen de puntos clave).

1. Observamos que las nuevas pruebas objetivas de personalidad se elaboran a un ritmo impresionante. Aparecen nuevos instrumentos en prácticamente cada número de revistas como *Psychological Assessment* o el *Journal of Personality Assessment*. También, de manera continua, se refinan los instrumentos existentes. Las editoriales producen en serie nuevas pruebas de personalidad y revisiones de pruebas antiguas a pasos vertiginosos. Concluimos que el campo de las pruebas objetivas de personalidad se encuentra rozagante de salud. Al parecer, los psicólogos ven una necesidad real de este tipo de herramientas. Por cierto, este rápido ritmo de desarrollo refuerza la necesidad de conocer las fuentes de información que tratamos en el capítulo 2.

2. Los métodos para elaborar pruebas objetivas de personalidad han madurado. Hemos recorrido un largo camino desde los primeros intentos crudos por medir los rasgos de personalidad y detectar padecimientos patológicos con inventarios de calificación objetiva. En particular, las influencias potenciales de la dirección y el falseamiento de respuestas se han reconocido ampliamente, y los métodos para tratar con estas influencias ya se han aplicado al proceso de elaboración de pruebas. Además, nos damos cuenta de la necesidad de demostrar la validez de la prueba de manera empírica en vez de basarnos en corazonadas acerca de qué mide la prueba. El uso del criterio meta y la investigación de las diferencias grupales son los métodos primarios para hacer tal demostración.

3. Uno de los contrastes clásicos al considerar los méritos relativos de diferentes tipos de pruebas es entre las pruebas objetivas de personalidad y las técnicas proyectivas, categoría que revisaremos en el capítulo 14. El contraste, por lo general, se hace en términos de confiabilidad y validez. Sin embargo, con el surgimiento del método de la atención administrada para contener los costos de los servicios de salud, el menor costo de las pruebas objetivas de personalidad, en comparación con las proyectivas, tiene cada vez más pertinencia (Piotrowski, 1999). Aunque un examinado puede dedicar casi la misma cantidad de tiempo a contestar una prueba proyectiva u objetiva, existe una diferencia considerable en el tiempo del clínico. Puede tomarle una hora aplicar una prueba proyectiva y otra hora calificarla. En el caso de una prueba objetiva de personalidad, el examinado puede registrar las respuestas en un teclado, de donde se transmiten a una computadora distante y, en cuestión de segundos, el clínico tiene el informe completo de las puntuaciones –sin haber dedicado tiempo a la calificación. Esta ventaja y el ahorro de costos tendrán, quizá, una importante influencia en el uso relativo de las pruebas proyectivas y objetivas. En particular, el ambiente de la

atención administrada favorece las pruebas objetivas y las de dominio específico más cortas en comparación con las técnicas proyectivas y los inventarios integrales, respectivamente.

4. Ahora son comunes los informes interpretativos del desempeño en pruebas objetivas de personalidad. Es importante que el estudiante reconozca las ventajas y los potenciales inconvenientes de tales informes. En el lado positivo, pueden ofrecer una interpretabilidad mejorada para el psicólogo y el cliente. Por otro lado, tenemos que tomar estos informes con las debidas reservas, pues pueden sonar demasiado definitivos. Tenemos que recordar que se basan en puntuaciones falibles (con una confiabilidad que no es perfecta) y que, por lo común, no incorporan información externa a la prueba misma. La responsabilidad última de la interpretación de la prueba pertenece al psicólogo que la usa, no al programa de cómputo que genera el informe.

5. El método tradicional para contestar una prueba objetiva de personalidad es el de lápiz y papel: sentarse y llenar la forma, en el cuadernillo de la prueba o en una hoja de respuestas separada. Cada vez más, la aplicación de estas pruebas se lleva a cabo en línea: sentarse ante una computadora o con el teléfono inteligente y responder las preguntas. Hasta ahora, estas aplicaciones sólo presentan la versión de lápiz y papel en forma electrónica, no en un modo adaptado para computadora, pero éste será más frecuente en los próximos años, como ya lo es en las pruebas de capacidad mental.

Resumen

1. La característica que define a un inventario objetivo de personalidad es el uso del formato de respuesta cerrada. También tiene, por lo general, afirmaciones cortas como troncos de los reactivos.
2. Las pruebas objetivas de personalidad tienen aplicaciones prácticas en una amplia variedad de áreas, como orientación, selección de personal e investigación.
3. Las pruebas objetivas de personalidad pueden clasificarse de manera conveniente en inventarios integrales y pruebas de dominio específico. Dentro de estas categorías, algunas pruebas se centran en rasgos normales de personalidad y otras en características anormales, en especial padecimientos patológicos.
4. La dirección y el falseamiento de respuestas pueden influir en las respuestas a las pruebas objetivas de personalidad. Se usan cuatro métodos para controlar o mitigar la influencia de estos factores: revisar las frecuencias empíricas extremas y la consistencia de las respuestas a reactivos similares, equilibrar la direccionalidad de los reactivos y usar reactivos de elección forzada.
5. Existen cuatro métodos para elaborar pruebas objetivas de personalidad: contenido, teórico, análisis factorial y criterio meta. Cada uno tiene sus ventajas y desventajas distintivas. En la práctica se puede emplear de manera habitual alguna combinación de estos métodos.
6. Estas pruebas ilustraron los inventarios integrales: el *Edwards Personal Preference Schedule*, como ejemplo del método teórico, y el NEO PI, como ejemplo del método analítico-factorial.
7. En el caso de las pruebas de dominio específico, se utilizó como ejemplo el *Pier-Harris Children's Self-Concept Scale* y una gran cantidad de escalas usadas en el recién surgido campo de la psicología positiva, que incluye medidas simples y multidimensionales del bienestar subjetivo.
8. Observamos ciertas tendencias en el uso actual de las pruebas objetivas de personalidad y especulamos acerca de su futuro en campos aplicados de la psicología. Sigue apareciendo un número impresionante de pruebas nuevas de esta categoría. La elaboración de pruebas ahora es más sofisticada. En la actualidad se usan más los informes interpretativos. La aplicación en línea ha aumentado su popularidad.

Palabras clave

análisis factorial
consistencia de las respuestas
criterio meta
dirección de las respuestas
distorsión de las respuestas
EPPS
estilo de respuesta
falseamiento negativo
falseamiento positivo
fingimiento
forense
frecuencia empírica extrema
inventario
inventario integral
manejo de impresiones
NEO PI
Piers-Harris
prueba de dominio específico
prueba estructurada
prueba objetiva de personalidad
psicología positiva
puntuaciones ipsativas
respuestas socialmente deseables

Ejercicios

1. Entra a la página de ETS Test Collection (http://www.ets.org/test_link/find_tests/). Escribe como palabra clave el nombre de algún rasgo de personalidad, el que te despierte un interés especial, como anxiety, depression, extroversion o *self-concept* (ansiedad, depresión, extroversión y autoconcepto, respectivamente). Observa la lista de pruebas que aparecen y trata de clasificar algunas en el esquema organizacional usado en el cuadro 12-3.

2. Varios inventarios de personalidad tienen su origen en el estudio sistemático de los *adjetivos* que usamos para describir las características de personalidad. Intenta hacerlo tú mismo. Enumera 20 adjetivos que describan la personalidad. Puede ser de ayuda pensar en cómo describirías a las personas que conoces. Compara tu lista con la de un compañero.

3. Recuerda que una debilidad potencial del método analítico-factorial para elaborar pruebas consiste en que si no se incluyen reactivos acerca de cierto rasgo en el fondo original de reactivos, no será posible identificar ese rasgo en el análisis factorial. Examina los reactivos del cuadro 12-2. ¿Qué rasgos de personalidad no parecen estar representados en estos reactivos? (Si hiciste el ejercicio 2, algunos de tus adjetivos pueden sugerir rasgos que se pasaron por alto en el cuadro 12-2).

4. Entra a la página de internet de la editorial de un inventario integral de los que aparecen en el cuadro 12-4 y encuentra la información acerca del inventario. ¿Qué características de la prueba destaca la página de la editorial? Busca en el apéndice C las direcciones de las editoriales.

5. Aquí se muestran respuestas de tres examinados a cuatro reactivos que podrían formar parte de un inventario de personalidad. ¿Qué respuestas parecen inconsistentes? Las respuestas son V = Verdadero y F = Falso.

	Examinado		
Reactivo	**Annika**	**Grazia**	**Olivia**
1. Me agrada la mayoría de personas.	V	F	V
2. Por lo general, trabajo duro.	F	F	F
3. Muchas personas son un fastidio.	F	V	F
4. En esencia, soy perezoso	V	V	F

Respuestas inconsistentes: Examinado _____ en los reactivos _____ y_____.

6. Supón que intentamos "falsear positivamente" para crear una impresión favorable. ¿Cómo responderías a los reactivos del ejercicio 5? Las respuestas son V = Verdadero y F = Falso.

7. Aquí se muestran las puntuaciones T de una persona en los cinco dominios del *NEO PI*. Para empezar el informe interpretativo, escribe tres oraciones que describan a esta

persona sin usar los nombres de los dominios.

Dominio: Neuroticismo Extroversión Franqueza Escrupulosidad Simpatía

Puntuación T: 60 45 40 65 45

8. Para contestar de manera gratuita un inventario parecido al NEO y obtener tu informe interpretativo, visita esta dirección: http://www.personal.psu.edu/j5j/IPIP/. Sigue las instrucciones que ahí se te indican.

CAPÍTULO 13
Instrumentos y métodos clínicos

Objetivos

1. Identificar semejanzas y diferencias entre pruebas de rasgos normales de personalidad e instrumentos clínicos.
2. Contrastar técnicas de entrevista clínica estructurada y no estructurada.
3. Describir las principales características de los instrumentos clínicos integrales, por ejemplo, MMPI-2, Millon y SCL-90-R.
4. Describir las principales características de los instrumentos clínicos de dominio específico, por ejemplo, BDI-II, EDI-3 y STAI.
5. Bosquejar el método usado para las escalas de valoración conductual.
6. Describir la noción básica de técnicas de evaluación conductual y dar ejemplos de aplicaciones específicas.
7. Discutir las tendencias en la elaboración y uso de instrumentos clínicos.

Introducción

En este capítulo se tratan varias pruebas y técnicas que se usan en contextos clínicos, como hospitales, práctica privada, centros de orientación y clínicas psicológicas en las escuelas. Las pruebas y técnicas que se abordan aquí tienen algunas semejanzas con las pruebas de rasgos normales de personalidad que revisamos en el capítulo anterior, pero también hay algunas diferencias (véase Resumen de puntos clave 13-1). Bosquejemos estas semejanzas y diferencias.

Quizá, la semejanza más evidente entre las dos categorías atañe a la naturaleza de los reactivos y los formatos de respuesta. Recordemos la descripción de estas dos categorías en el capítulo anterior. Las pruebas tienden a usar afirmaciones muy sencillas como reactivos, por ejemplo, "A menudo estoy triste", y formatos de respuesta también muy sencillos, por ejemplo, "Sí-No" o "De acuerdo-No estoy seguro-En desacuerdo". La segunda semejanza, introducida en el capítulo anterior, es que las dos categorías están convenientemente subdivididas en instrumentos integrales y de dominio específico, esquema organizacional que empleamos en este capítulo. Los instrumentos clínicos integrales intentan investigar todas las áreas potenciales de dificultad, por lo que producen numerosas puntuaciones. En cambio, los instrumentos de dominio específico se concentran en sólo un área, por ejemplo, depresión o trastornos de la conducta alimentaria, lo cual produce sólo una o algunas puntuaciones estrechamente relacionadas. Las pruebas de ambas categorías se han elaborado con estrategias similares, pero el criterio meta ha sido más importante en el caso de los instrumentos clínicos. También comparten la preocupación por la dirección y el falseamiento de respuestas.

Existen varias diferencias entre las categorías de pruebas tratadas en éste y en el capítulo anterior. La más evidente es su orientación; las pruebas del capítulo anterior se relacionan primordialmente con el rango normal de los rasgos de personalidad, mientras que las pruebas y otras técnicas de evaluación que se revisan en este capítulo se relacionan primordialmente con la psicopatología o, al menos, con cierta dificultad personal. Hay algunas diferencias menos evidentes entre las dos categorías. Los instrumentos clínicos casi siempre se aplican de manera individual, sobre todo en el contexto clínico, mientras que las pruebas del capítulo anterior a menudo se aplican en contextos grupales. Los manuales de los instrumentos clínicos hacen hincapié en el uso de los resultados para diagnosticar, planear el tratamiento y hacer una evaluación de seguimiento. No encontramos estas discusiones en los manuales de pruebas como el NEO PI-R.

Aunque tratamos pruebas de rasgos normales de personalidad e instrumentos clínicos como categorías separadas, debemos admitir que se superponen en cierta medida entre sí. Por ejemplo, el *Piers-Harris Children's Self-Concept Scale*, revisado en el capítulo anterior, podría usarse para explorar estudiantes con un autoconcepto muy bajo, lo que sugeriría la necesidad de una evaluación más profunda. Al mismo tiempo, el *Eating Disorder Inventory*, que revisaremos en este capítulo, tiene su propia escala de

autoestima, con reactivos muy similares a los del Piers-Harris. Así, la distinción que usamos en estos dos capítulos no es irrefutable.

Por último, señalaremos que las pruebas del siguiente capítulo, técnicas proyectivas, se usan casi exclusivamente como instrumentos clínicos. Sin embargo, son diferentes en su estructura con respecto de los instrumentos que veremos en este capítulo, lo que justifica su revisión por separado.

Resumen de puntos clave 13-1

Comparación de pruebas de los capítulos 12 y 13

Semejanzas

- Naturaleza de los reactivos y los formatos de respuesta
- Subdivisiones en pruebas integrales y de dominio específico
- Estrategias de elaboración
- Preocupación por la dirección y el falseamiento de las respuestas

Diferencias

- Orientación
- Escenarios de aplicación
- Diagnóstico, planeación del tratamiento y seguimiento

Entrevista clínica como técnica de evaluación

Al introducir las distintas pruebas a lo largo de este libro, nos hemos referido con frecuencia a la clasificación de las pruebas en los reportes sobre su uso, en los cuales nos hemos basado para decidir qué pruebas describir. De acuerdo con el énfasis del libro en el aspecto práctico, nos hemos concentrado en las pruebas más usadas. Cuando incluimos "entrevista" en las encuestas de "más usadas" siempre se ubica en el primer lugar (véase, p. ej., Culross & Nelson, 1997; Watkins, Campbell, Nieberding, & Hallmark, 1995). Desde luego, incluir "entrevista" como si se tratara de una prueba específica, como el WAIS o el MMPI, es engañoso. Es como preguntar: ¿alguna vez obtienes información biográfica, como la edad y el sexo, de tus clientes? Bueno, por supuesto, pero eso no es una prueba. Tampoco la entrevista es una prueba única y específica, por lo que no se puede comparar con otras pruebas. No obstante, los resultados de encuestas nos recuerdan que la entrevista clínica se usa casi de manera universal y también que necesitamos hacernos las mismas preguntas acerca de la entrevista, como si se tratara de cualquier otra prueba: ¿Es confiable? ¿Es válida? ¿Es neutral? ¿Es viable en los costos que genera?

La investigación sobre lo que podría llamarse entrevista tradicional ha producido respuestas dolorosamente pesimistas; es decir, la entrevista tradicional no es muy confiable, tiene una validez limitada, es susceptible de contener sesgos y no es muy viable en términos de su costo. En Groth-Marnat (2009) se puede encontrar un resumen útil de la investigación sobre estas preguntas. La preocupación por lo adecuado del aspecto técnico de la entrevista clínica tradicional, aumentada por las demandas de la atención administrada, ha llevado al desarrollo de la **entrevista clínica estructurada**, el principal tema de esta sección.

Entrevistas estructuradas, semiestructuradas y no estructuradas

En el campo de las entrevistas clínicas es habitual distinguir entre entrevistas estructuradas, semiestructuradas y no estructuradas. La *entrevista no estructurada*, a veces llamada entrevista "tradicional", no sigue un patrón particular y varía de un cliente a otro, así como de un clínico a otro, lo cual no significa que sea irreflexiva o descuidada. En el otro extremo, la entrevista clínica estructurada busca abordar los mismos temas con las mismas preguntas con todos los clientes. Aspira a ser exhaustiva y consistente; dentro de la consistencia se consideran las preguntas y la codificación de las respuestas. Además, existe un registro específico de respuestas, no sólo una impresión o un recuerdo general. El método semiestructurado, como podrás adivinar, se ubica entre los métodos estructurado y no estructurado. Tiene algunas preguntas estándar, pero está hecho a la medida del cliente.

Estos tres métodos no son categorías discretas, sino que forman parte de un continuo que va desde la total unicidad hasta la total igualdad. Por ejemplo, si es evidente que un

cliente está lúcido pero se presenta con quejas de disforia, no tiene mucho sentido dedicar una gran cantidad de tiempo a indagar sobre alucinaciones, incluso si estas preguntas son parte de un inventario de entrevista completa estructurada.

El DSM y el CIE

No es posible comprender la mayoría de las entrevistas estructuradas contemporáneas sin hacer referencia a alguna versión del Manual Diagnóstico y Estadístico de los Trastornos Mentales (**DSM**, siglas en inglés; *American Psychiatric Association*, 2000, 2013). Sin duda, presentar una descripción razonable del DSM nos alejaría demasiado de nuestro principal interés; sin embargo, debemos señalar las siguientes características de esta fuente crucial. Primero, es el estándar común de las categorías diagnósticas en la práctica clínica contemporánea, por lo que ha cobrado una enorme importancia para la comunicación entre profesionales. También se ha vuelto esencial para la mayoría de demandas por reembolso. Segundo, el principal resultado de usar el DSM es llegar a una categoría diagnóstica, en esencia una etiqueta, por ejemplo, 309.21 trastorno de ansiedad de separación. Así, da por resultado una clasificación nominal, aunque algunas categorías tienen subdivisiones de acuerdo con distintos niveles de gravedad. Tercero, se orienta a los síntomas; es decir, la tarea clave del clínico es identificar los síntomas que se enumeran en el DSM relacionados con un padecimiento. Estas tres características ofrecen la base de la mayoría de las entrevistas clínicas estructuradas: determinar los síntomas y llegar a una clasificación del DSM.

Ha coincidido con la publicación de la versión en inglés de este libro la quinta edición del DSM en inglés: DSM-5. Observa la loca carrera entre editoriales y autores de pruebas actuales para mostrar que actualizan sus pruebas de acuerdo con el DSM-5, quizá haciendo pequeñas modificaciones a sus pruebas. También observa la aparición de nuevas pruebas diseñadas tomando como base el DSM-5.

La principal alternativa al DSM es la Clasificación Internacional de Enfermedades (**CIE**) financiada por la Organización Mundial de la Salud (OMS). El CIE se encuentra en su décima edición, a la que se hace referencia como CIE-10; y la edición 11 se espera para 2015. Por órdenes de agencias federales, los proveedores de servicios de salud, entre ellos psicólogos, deben adoptar el sistema del CIE a partir de octubre de 2014 para los servicios de reembolso. Por lo tanto, es probable que veamos en el futuro inmediato que los manuales de las pruebas se orientan hacia el CIE en vez del DSM.

Entrevista Clínica Estructurada del DSM-IV para los Trastornos del Eje I

La *Structured Clinical Interview for DSM-IV Axis I Disorders* [Entrevista clínica estructurada del DSM-IV para los trastornos del Eje I] (**SCID-I**; First, Spitzer, Gibbon, & Williams, 1997) es la entrevista clínica estructurada a la que más se hace referencia. Stewart y Williamson (2004) afirmaron que "el SCID es el 'estándar de oro' para el

diagnóstico objetivo y válido de trastornos psiquiátricos comórbidos" (p. 186). Es interesante que la guía del usuario del SCID también emplee la expresión estándar de oro, pero lo hace en un sentido negativo, pues señala que "desafortunadamente, un estándar de oro para el diagnóstico psiquiátrico sigue siendo escurridizo" (First *et al*., 1997, p. 46). Así, según parece, nos encontramos en la delicada posición de tener un estándar de oro que no está relacionado con el oro. Es decir, en la medida en que el sistema del DSM es defectuoso o ambiguo, un instrumento apegado específicamente a dicho sistema casi sin duda tendrá los mismos problemas. No obstante, aquí presentamos una breve descripción de la SCID.

Pero antes de hacerlo, serán útiles algunas palabras sobre la "jerga" de la entrevista clínica estructurada. Aunque el instrumento que describimos se denomina "entrevista clínica estructurada", muchos instrumentos de este tipo incorporan el término entrevista clínica estructurada en sus nombres. Rogers (2001) y Groth-Marnat (2009) hacen excelentes descripciones de una docena de estos instrumentos. Además, existen varias versiones del instrumento que describimos aquí; por ejemplo, la SCID-I tiene una versión clínica y una versión de investigación más larga. También tenemos una SCID basada en el DSM-III y otra, en el DSM-IV; hay una SCID para los trastornos del eje I y otra para los del eje II. Por último, en la práctica, un clínico que usa la SCID (cualquier versión o cualquier otra entrevista clínica estructurada), por lo general, aplica sólo una parte del instrumento. Así, referirnos al uso de la entrevista clínica estructurada o, incluso, a la SCID tiene un sentido ambiguo.

¡Inténtalo!

Visita el sitio www.scid4.org para ver la gran cantidad de versiones actuales de la SCID y consulta resúmenes acerca de su confiabilidad y validez.

La SCID-I-CV funciona de la siguiente manera:

- Hay un cuadernillo de aplicación que contiene preguntas e instrucciones para proceder, y una hoja de calificaciones para registrar y codificar las respuestas. El entrevistador los tiene abiertos al mismo tiempo.
- Después de registrar la información de identificación básica, se comienza con las principales categorías del DSM, por ejemplo, trastornos del estado de ánimo o cualquier otro que parezca apropiado. Estas categorías corresponden a los módulos de la SCID, que se muestran en el cuadro 13-1.
- Se hacen preguntas acerca de los síntomas, como lo indica el cuadernillo de aplicación.
- Se registran las respuestas, pero también se codifican señalando presencia (+), ausencia (−) o información inadecuada (?) en cada respuesta.
- Se observa si se siguen las reglas (véase más adelante).
- Cuando la entrevista concluye, se llena el resumen diagnóstico, el cual relaciona

las respuestas con los números de categorías y etiquetas del DSM.

Cuadro 13-1. Lista de módulos del SCID-I-Clinical Version

A. Episodios del estado de ánimo	D. Trastornos del estado de ánimo
B. Trastornos psicóticos	E. Trastornos por abuso de sustancias
C. Síntomas psicóticos	F. Trastornos de ansiedad y otros relacionados

El código "ausencia" se aplica no sólo para referirse a la ausencia absoluta del síntoma, sino también a los llamados niveles de "subumbrales", es decir, algo que no supera el nivel indicado como clínicamente significativo en el DSM. No seguir las reglas lleva al entrevistador a dejar de hacer preguntas de una categoría particular cuando es evidente que no es de provecho continuar por esa línea. Estas reglas funcionan como un "alto" en las pruebas de inteligencia de aplicación individual. En una aplicación típica, en la que sólo se puede usar uno o algunos módulos, contestar la SCID lleva cerca de una hora, mientras que contestar todos los módulos, que suele hacerse sólo en las aplicaciones de investigación, puede llevar tres horas.

La SCID intenta ser una opción intermedia entre la estandarización completa y rígida y la entrevista tradicional no estructurada, pero es claro que se inclina más hacia el lado estandarizado. Por ejemplo, en la lista de qué hacer y no hacer que aparece en la guía del usuario, encontramos: "Siga las preguntas iniciales tal y como están escritas, a excepción de modificaciones menores" y "No invente preguntas iniciales por sentir que usted tiene una mejor manera de obtener la misma información". Por otro lado, "Utilice su criterio acerca de un síntoma tomando en cuenta toda la información disponible" y "No se limite a obtener una respuesta sólo mediante las preguntas de la SCID-CV" (First *et al*., 1997, pp. 12-13).

Por último, debe hacerse hincapié en que la entrevista clínica no tiene el propósito de ser la única fuente de información sobre el cliente, sino que puede complementarse con varias pruebas como las que presentamos en otros capítulos de este libro, por ejemplo, el MMPI-2, BDI-II o alguna prueba neuropsicológica. También se puede obtener información de los miembros de la familia; en algunas circunstancias se puede ordenar un examen médico.

Entrevista de trabajo

Nos hemos concentrado en la entrevista clínica; por lo común, un psicólogo clínico, un psiquiatra, un trabajador social u otro profesional la lleva a cabo en la visita inicial del cliente. La meta principal es el diagnóstico, seguido de la planeación del tratamiento. Sin embargo, prácticamente todo lo que se ha dicho de la entrevista clínica se aplica igual de bien a la entrevista de trabajo, sólo que la meta es la selección de un empleado en vez del diagnóstico. En particular, la investigación sobre la entrevista de trabajo muestra que la

entrevista típica tiene una confiabilidad y validez notablemente bajas. En el caso de la entrevista clínica, el remedio que se ha propuesto es la entrevista estructurada. En Huffcutt (2011), Huffcutt y Youngcourt (2007) y Macan (2009) se puede encontrar revisiones de la investigación sobre las entrevistas de trabajo. De manera paralela a los criterios de especificidad y pertinencia usados en la entrevista clínica estructurada, la entrevista de trabajo estructurada hace hincapié en a) hacer las mismas preguntas a todos los candidatos, b) codificar las respuestas y c) concentrarse en las áreas directamente pertinentes para el puesto.

¡Inténtalo!

Supón que estás elaborando una entrevista estructurada para contratar maestros de universidad. ¿Qué preguntas harías a los aspirantes? Asegúrate de que las preguntas se relacionen directamente con las habilidades que se requieren para el puesto.

Ejemplos de inventarios integrales de autoinforme

Igual que en el capítulo anterior, presentamos ejemplos de instrumentos integrales y de dominio específico con una orientación clínica. Las entrevistas estructuradas de las que hablamos fueron, por lo común, integrales, pero las distinguimos por su naturaleza. Ahora nos referiremos estrictamente a los autoinformes.

Inventario Multifásico de Personalidad de Minnesota (MMPI)

Una elección fácil –aunque también se podría decir obligada– como ejemplo de un instrumento clínico integral de autoinforme es el Inventario Multifásico de Personalidad de Minnesota (**MMPI**), conocido en su segunda edición, la actual, como MMPI-2. Con 567 reactivos, el MMPI-2 requiere de 60 a 90 minutos para contestarlo y puede llegar a tomar hasta dos horas a examinados con niveles bajos de lectura o niveles altos de distraibilidad. Este instrumento está en casi cualquier lista de los más usados, investigados y citados. Por ejemplo, es la prueba más usada por los neuropsicólogos y la segunda más usada por los psicólogos clínicos (Camara *et al*., 2000). Se ha empleado en más de 10 000 estudios publicados. Además de la frecuencia con que se usa, el MMPI fue pionero en el desarrollo de índices de validez y del método de criterio meta en la elaboración de pruebas. De ahí que sea una fuente de extraordinaria utilidad para el estudiante de psicometría.

El uso del MMPI es prácticamente una subcultura dentro de la psicología. Tiene su propio lenguaje, costumbres y rituales. Para los iniciados en esta subcultura, es rico en tradición y significado; por ejemplo, se puede caracterizar a un examinado como "tipo código 24 con una escala F elevada". Esta breve descripción le dice mucho al clínico conocedor del MMPI. Para los no iniciados, puede ser incomprensible. La caracterización de una persona es galimatías para alguien que no conoce el funcionamiento interno del MMPI. En el poco espacio de que disponemos, sólo podemos presentar los puntos más importantes de la estructura de la prueba, su uso y la investigación en que se fundamenta. El lector interesado en esta prueba puede consultar Graham (2006), Newmark y McCord (1996), las reseñas estándar (Archer, 1992; Duckworth & Levitt, 1994; Nichols, 1992) y, desde luego, los manuales del MMPI-2 (Butcher, 1993; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989).

Primera edición

El MMPI se publicó por primera vez en 1942, mientras que la edición revisada, MMPI-2, que es nuestro principal interés, apareció en 1989. Sin embargo, el MMPI es una de

las grandes sagas de la psicometría, por lo que debemos dedicar más tiempo a describir la primera edición y el contexto en que se elaboró.

Cuando se estaba elaborando, en la década de 1930, la evaluación de la personalidad estaba dominada por los inventarios basados en el contenido, los cuales estaban plagados de problemas de dirección y falseamiento de respuestas. Además, se cuestionaba el significado de sus puntuaciones, en especial para el uso clínico. Haciendo a un lado los problemas de la dirección de respuestas, ¿qué decían las puntuaciones al clínico? Desde luego, el significado para el clínico estaba definido en gran medida por lo que el clínico trataba de hacer. En ese tiempo, una preocupación importante era el diagnóstico exacto de los trastornos psicológicos, que resultaba en la aplicación de la categoría diagnóstica correcta (en la terminología vigente en aquel tiempo) a un caso. El MMPI original tenía dos características clave. Primero, los índices de validez se usaban de manera explícita (y se describen más adelante junto con otras puntuaciones del MMPI). Segundo, la prueba usaba el método de criterio meta para desarrollar nueve escalas clínicas; después se agregó otra escala para elevar el número a 10. Los reactivos de cada escala (excepto Introversión social, Si) se eligieron contrastando respuestas de casos claramente diagnosticados en la categoría con un grupo de personas "normales", que fueron 724 visitantes del hospital. La selección de reactivos para las escalas clínicas también se sometió a una validación cruzada. Toda esta investigación (nuevamente, a excepción de la escala Si) se llevó a cabo en el hospital de la Universidad de Minnesota. Las personas normales sirvieron como base de las normas de puntuaciones T y percentiles del MMPI. La prueba fue criticada de manera consistente por su base normativa restringida: un grupo relativamente pequeño de individuos, del mismo estado, casi todos blancos, primordialmente rurales y, obviamente, limitados a los últimos años de la década de 1930. La décima escala (etiquetada como 0 en vez de 10) se elaboró después en la Universidad de Wisconsin.

El MMPI rápidamente se convirtió en el instrumento preferido de los clínicos para la evaluación objetiva de la personalidad. Se usó mucho no sólo en la práctica sino en la investigación. Conforme esta última creció, la interpretabilidad de las puntuaciones mejoró. De hecho, el MMPI es quizá el mejor ejemplo de cómo la utilidad práctica de una prueba psicológica aumenta después de ser publicada gracias a la investigación.

¡Inténtalo!

Para apreciar qué tanto se usa esta prueba en la investigación, escribe MMPI-2 como palabra clave en una base de datos electrónica como PsychINFO. Quizá sea mejor limitar la búsqueda a sólo algunos años. Examina los estudios que aparecen y determina cómo se usa la prueba.

La revisión de 1989

Debido al uso extendido y a la considerable investigación sobre el MMPI original, la

revisión se realizó con cierto temor. En realidad, las revisiones fueron bastante modestas, por lo que la investigación sobre la comparabilidad de la primera y segunda ediciones sugiere una transición razonablemente suave (Graham, 2006). Hubo cinco principales tipos de revisión.

Primero, algunos reactivos fueron revisados o reemplazados, en especial los que contenían referencias obsoletas o específicas al género. Segundo, las escalas clínicas se citaron mediante un número (1, 2, 3...) en vez de la categoría diagnóstica. Tercero, se crearon normas por completo nuevas (la reestandarización se describe más adelante). Cuarto, se agregaron varios índices de validez nuevos. Por último, para considerar alta una puntuación, la puntuación estándar se redujo de 70 a 65.

Resumen de puntos clave 13-2

Principales categorías de puntuaciones del MMPI-2
Índices de validez
Escalas clínicas
Escalas de contenido
Escalas complementarias
Reactivos críticos
Tipos de código

Puntuaciones

El MMPI-2 es en verdad abundante en puntuaciones. Por lo general, podemos decir de manera definitiva que una prueba produce, digamos, 5 o 12 puntuaciones, pero en el caso del MMPI-2 es difícil decir con exactitud el número de puntuaciones que produce. Intentemos revisar al menos las principales seis categorías de puntuaciones de esta prueba (véase Resumen de puntosclave 13.2).

Primero, están los índices de validez, entre los que se encuentran cuatro índices tradicionales que aparecieron en el MMPI original y se mantuvieron en el MMPI-2. Además, hay tres índices nuevos. Existe una amplia investigación en cuanto a los índices tradicionales, que se informan de manera habitual en el perfil básico del MMPI-2 (véase figura 13-1). Las puntuaciones de los índices de validez nuevos no aparecen de manera habitual en el perfil básico del MMPI-2. El cuadro 13-2 ofrece una breve descripción de cada índice de validez.

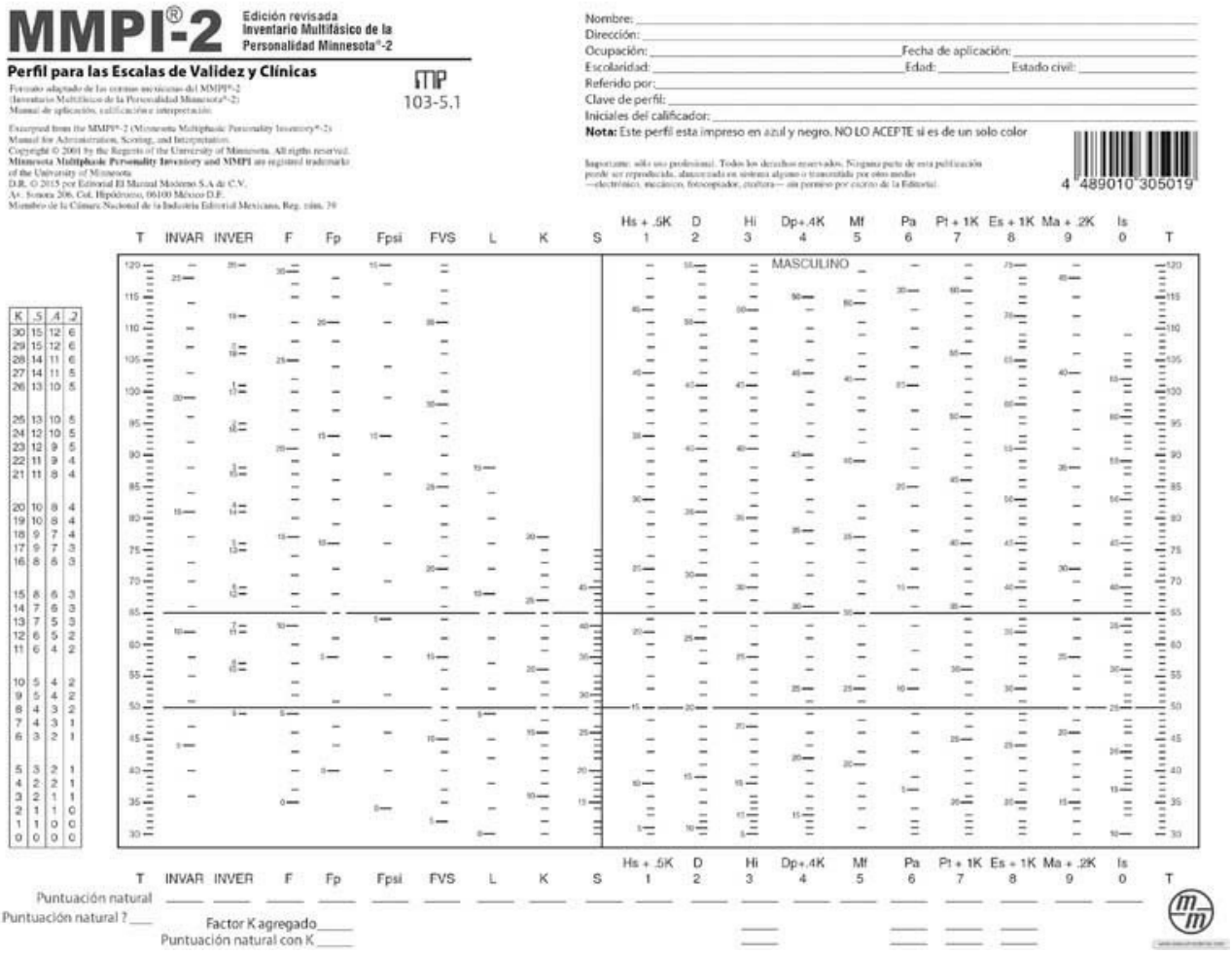


Figura 13-1. Muestra del Perfil para las Escalas de Validez y Clínicas del MMPI-2 edición revisada.

Fuente: MMPI®-2 Edición revisada Inventario Multifásico de la Personalidad Minnesota®-2.

D. R. © 2015 por Editorial El Manual Moderno S. A. de C. V.

Cuadro 13-2. Índices de validez del MMPI-2

Abreviatura	Nombre	Reactivos	Breve descripción
Índices tradicionales			
?	No puedo decir	Todos	Simple conteo del número de reactivos omitidos o con doble marca.
L	Mentira	15	Las respuestas pueden indicar una tendencia a crear una impresión favorable o a fingir una buena imagen.
F	Infrecuencia	60	Reactivos con una baja tasa de aprobación. Las respuestas pueden indicar el intento de fingir una mala imagen.
K	Corrección	30	Las respuestas pueden indicar una tendencia a fingir una buena

			imagen, pero en un nivel más sutil que el que indican los reactivos de la escala L.
Nuevos índices			
TRIN	Inconsistencia en las respuestas "verdadero"	23 pares	Número de respuestas "verdadero" (o "falso") en los pares de reactivos con contenido opuesto. Es inconsistente responder "verdadero" (o "falso") en los dos reactivos de un par.
VRIN	Inconsistencia en las respuestas variables	67 pares	Número de respuestas inconsistentes a los pares que tienen un contenido similar u opuesto.
F_B	Infrecuencia posterior	40	Los reactivos tipo F que se presentan al final de la prueba ayudan a determinar si los patrones finales de respuesta son similares a los iniciales.

Hay una buena razón para empezar la descripción del MMPI-2 con los índices de validez. En la práctica, la interpretación de todas las puntuaciones del MMPI-2 comienza con el examen de estos índices. Si uno o más es extremadamente atípico, puede ser que las demás puntuaciones no se interpreten, pues podemos concluir que el examinado respondió de manera aleatoria o intentó dar una buena o una mala impresión, las respuestas, de modo que la prueba entera es inútil. En casos menos graves, estos índices pueden sugerir que se debe tener un cuidado excepcional al interpretar las puntuaciones.

Debe hacerse una mención especial de la puntuación K (Corrección). No sólo es una puntuación en sí misma, sino que lleva a modificar varias puntuaciones de las escalas clínicas. En la figura 13-1, podemos observar la entrada en la parte inferior del perfil: "K a agregarse". También podemos observar el estrecho cuadro etiquetado como "Fracción de K" a la izquierda del perfil. Por último, observamos las etiquetas de las escalas clínicas en la parte inferior del perfil; algunas abreviaturas hacen referencia a K. Por ejemplo, la primera escala (Hs, 1) está etiquetada como Hs + .5K, lo cual significa que se debe agregar .5 × puntuación K a la puntuación original de Hs. Las escalas 7, 8 y 9 también tienen "correcciones K"; en efecto, los autores de la prueba nos están diciendo: vamos a inflar tu nivel informado de desviación mediante una cantidad proporcional a tu esfuerzo para falsear positivamente las respuestas con el fin de obtener un indicador más exacto de tu nivel real de desviación. Los autores concluyeron que esta compensación en las puntuaciones sólo estaba justificada en el caso de ciertas escalas.

La interpretación de los índices de validez no es tan simple como parece al principio. Por ejemplo, una puntuación F elevada, más que indicar el esfuerzo para falsear negativamente las respuestas, puede indicar una patología grave o "un grito de ayuda". Una puntuación F_B alta, basada en los reactivos de aproximadamente la última tercera parte de la prueba, puede indicar que la atención del examinado osciló o incluso que se tornó hostil al final de la sesión de evaluación. Sin embargo, puede no invalidar las escalas clínicas, porque sus reactivos están predominantemente en las primeras dos terceras partes de la prueba.

Las *escalas clínicas* constituyen la segunda categoría principal de puntuaciones. Son

las escalas más distintivas del MMPI-2 y las que dan a esta prueba su sabor especial. El cuadro 13-3 presenta las 10 escalas clínicas. Su origen en basado en el criterio meta se describió antes. En varios contextos, una de estas escalas se puede citar mediante su número, su nombre original o la abreviatura de su nombre. El número de reactivos varía de una escala a otra, de 32 a 78 reactivos, con una mediana de 49. Existe cierta superposición de los reactivos entre las distintas escalas, es decir, un reactivo puede formar parte de más de una escala. Las escalas clínicas y los índices de validez tradicionales constituyen lo que se denomina *escalas básicas* del MMPI-2 y son las que aparecen en el perfil de la figura 13-1.

Cuadro 13-3. Escalas clínicas del MMPI-2

Número de la escala	Nombre original	Abreviatura	Número de reactivos
1	Hipocondriasis	Hs	32
2	Depresión	D	57
3	Histeria	Hy	60
4	Desviación psicopática	Pd	50
5	Masculinidad-Feminidad	Mf	56
6	Paranoia	Pa	40
7	Psicastenia	Pt	48
8	Esquizofrenia	Sc	78
9	Hipomanía	Ma	46
0	Introversión social	Si	69

Las *escalas de contenido* conforman la tercera categoría de puntuaciones. Aunque el MMPI es mejor conocido por su método criterio meta para elaborar las escalas, las de contenido se elaboraron mediante el análisis racional de los reactivos. Clínicos expertos examinaron los reactivos y agruparon los que parecían medir el mismo constructo. Estos grupos de reactivos se refinaron más adelante mediante un minucioso examen de las correlaciones reactivo-total para mejorar la homogeneidad de la escala. Butcher *et al*. (1989) identifican 15 escalas de contenido: ansiedad, temores, obsesividad, depresión, preocupaciones por la salud, actividad mental extraña, ira, cinismo, prácticas antisociales, Conducta tipo A, baja autoestima, incomodidad social, problemas familiares, interferencia en el trabajo e indicadores negativos del tratamiento. El número de reactivos por escala varía de 16 a 33, con una mediana de 24. Así, las escalas de contenido son considerablemente más cortas, en promedio, que las escalas clínicas. Las puntuaciones T de las escalas de contenido, por lo regular, aparecen en un perfil diferente del de las escalas básicas en los informes.

¡Inténtalo!

Examina los reactivos del cuadro 12-2 e identifica subgrupos de ellos que pienses que se relacionan

con el mismo constructo o dominio de contenido.

La cuarta categoría de las puntuaciones del MMPI-2 constituye una abrumadora lista de *escalas complementarias*, quizá tantas que puede no haber una cuenta definitiva. Algunas de ellas se han construido con el método de criterio meta para grupos adicionales (p. ej., alcohólicos vs. normales). Otras escalas han surgido del análisis racional, similares a las escalas de contenido; otras más son resultado del análisis factorial del fondo de reactivos del MMPI-2. Algunas escalas complementarias han sido investigadas y se han incorporado a los informes del MMPI-2, aunque algunas veces sólo de manera tentativa. Otras no se han investigado tanto, por lo que suelen no informarse. No describiremos o enumeraremos estas escalas; sin embargo, el lector debe tener en mente su existencia.

La quinta categoría de puntuaciones consiste en lo que se denomina **reactivos críticos.** Estas "puntuaciones" son literalmente sólo informes de las respuestas a los reactivos individuales; sin embargo, dichos reactivos son los que parecen tener una importancia especial. Por ejemplo, supongamos que una prueba contiene el reactivo: A menudo tengo ganas de cortarme las venas. Sin importar la contribución de este reactivo a la puntuación total, una respuesta Verdadero captará nuestra atención y querremos explorar esta respuesta con el cliente. Hay varias listas de reactivos críticos que, por lo general, se identifican por el nombre de la persona que creó la lista. Igual que con las escalas complementarias, no catalogaremos estas listas, pero el lector debe tener en mente que los reactivos críticos constituyen una categoría en el informe.

La sexta y última categoría consta de *tipos de códigos*. Ya que éstos son un mecanismo para informar, así como una categoría de puntuaciones, los describiremos en la siguiente sección.

Normas e informes

Las normas del MMPI-2 se basan en 1138 hombres y 1462 mujeres, 2600 en total, provenientes de seis estados, varias bases militares y una reserva de indios. El manual de la prueba presenta un análisis detallado de la muestra de estandarización en términos de edad, estado civil, raza, ingresos, educación y ocupación. Esta muestra es razonablemente representativa de la población adulta de EUA en términos de la mayoría de las características, pero están sobrerrepresentados los niveles altos de educación y de ocupación profesional. Las puntuaciones naturales se convierten en puntuaciones T y percentiles. Las puntuaciones T son el medio típico con que se informan los resultados de la prueba como se ilustra en el perfil de muestra que aparece en la figura 13-1. Los percentiles son específicos del género. Las normas del MMPI-2 constituyen una mejora importante con respecto de las normas del MMPI original.

Existen tres métodos principales para informar las puntuaciones del MMPI-2: perfiles, narraciones y tipos de códigos. Primero, el perfil (figura 13-1) es un mecanismo común para el informe; en él aparecen los índices tradicionales de validez y las escalas clínicas. Un perfil configurado de manera similar se usa para las escalas de contenido.

El segundo método es una narración o un informe interpretativo. En realidad, existen varios tipos de estos informes, desarrollados y distribuidos por diferentes vendedores. Pearson Assessment, Inc., en la actualidad, ofrece la asombrosa cantidad de 60 diferentes informes interpretativos para el MMPI. Los informes, a menudo de 15 páginas, incorporan perfiles de los índices de validez y de las escalas clínicas, de contenido y complementarias, así como las respuestas a los reactivos críticos. La parte narrativa del informe trata cuestiones como la validez del perfil, patrones sintomáticos y consideraciones del tratamiento, entre otros temas.

¡Inténtalo!

Para ver ejemplos de los muchos informes interpretativos del MMPI, sólo escribe "MMPI interpretive reports" [informes interpretativos del MMPI] en cualquier buscador de internet y abre los vínculos.

Tipos de códigos «341-342a

El tercer método importante para informar las puntuaciones del MMPI-2 son los **tipos de códigos**. Este tipo de informe es casi exclusivo de este instrumento; en esencia, se trata de un nuevo tipo de puntuación normativa y, por lo tanto, requiere una introducción. Una idea interpretativa clave con el MMPI es el examen de los picos del perfil de puntuaciones; sin importar si las puntuaciones más altas son extremada o moderadamente altas, se le otorga un significado especial a las puntuaciones más altas del perfil. Consideremos el sencillo ejemplo del cuadro 13-4: en las pruebas A, B, C y D, Sue y Bill tienen puntuaciones muy diferentes. Sin embargo, cuando sus puntuaciones se clasifican en orden, dentro de sus propios perfiles, sus puntuaciones tienen el mismo orden; ambos tienen "picos" en C y D. En el esquema del MMPI-2, ellos serían clasificados como tipo de código CD. En las pruebas C y D, Anne tiene las mismas puntuaciones que Sue, pero Anne es un tipo de código diferente porque sus puntuaciones más altas son en A y B.

Cuadro 13-4. Ilustración del tipo de código para identificar las puntuaciones más altas de un perfil

Prueba	**A**	**B**	**C**	**D**				
	Puntuaciones T				**Puntuaciones en orden**			
Sue	60	55	80	75	C	D	A	B
Bill	55	45	65	60	C	D	A	B
Anne	85	90	80	75	B	A	C	D

Los tipos de código del MMPI-2 utilizan los números de las escalas clínicas y las letras de los índices tradicionales de validez L, F y K (cuadros 13-2 y 13-3). Los números de

las escalas están ordenados de la puntuación T mayor a la menor de izquierda a derecha, seguidos del ordenamiento de los índices de validez. Los símbolos (como ! y *) se usan después del número de una escala para indicar el nivel absoluto de la elevación de la puntuación. En Butcher *et al*. (1989) o Graham (2006) se puede encontrar un conjunto completo de reglas de codificación. En la práctica real se debe atender principalmente a las dos puntuaciones más altas; esto se denomina tipo de código de dos puntos, por ejemplo, un tipo de código 2–4 o 4–2. Las dos puntuaciones más altas, por lo general, se consideran intercambiables, de modo que 2–4 y 4–2 se consideran el mismo tipo de código. También hay un tipo de código de tres puntos, pero el sistema de dos puntos ha sido el más usado.

Se ha llevado a cabo una considerable cantidad de investigaciones para determinar las características de personas con ciertos tipos de códigos. Por ejemplo, Graham (2006, p. 100) informa que quienes tienen un tipo de código 24/42 "parecen estar enojados, resentidos, y son hostiles, críticos y les gusta discutir... [y] ... el pronóstico de la psicoterapia tradicional no es bueno". Conocer la investigación sobre los tipos de código puede brindar al clínico un marco interpretativo importante.

¡Inténtalo!

¿Cuál es el tipo de código de dos puntos del perfil del MMPI-2 que se muestra en la figura 13-1? Asegúrate de usar puntuaciones T y no las puntuaciones naturales para determinar el tipo de código.

Confiabilidad y validez

El cuadro 13-5 resume los datos de confiabilidad del manual del MMPI-2 (Butcher *et al*., 1989). Los datos incluyen rangos y medianas de los coeficientes de consistencia interna y de confiabilidad de test-retest. El manual presenta todos los datos separados por género, pero el resumen que se presenta aquí reúne los datos de hombres y mujeres. En el manual aparecen datos adicionales de algunas escalas complementarias, pero no están incluidos en nuestro resumen.

Cuadro 13-5. Resumen de los datos de confiabilidad de las puntuaciones del MMPI-2

Escalas	Alpha		Test-retest	
	Mediana	Rango	Mediana	Rango
De validez	.64	.57-.73	.80	.69-.84
Clínicas	.62	.34-.87	.82	.58-.92
De contenido	.86	.67-.86	.85	.78-.91

Los datos del cuadro 13-5 sugieren varias conclusiones. Primero, la consistencia interna de los índices de validez y las escalas clínicas es, por lo general, débil; algunas escalas tienen una confiabilidad de consistencia interna sumamente baja. En el caso de

las escalas clínicas, aunque quizá esto es comprensible por el método con que se elaboraron, también es desconcertante. La confiabilidad de test-retest es notablemente mejor en las escalas de validez y clínicas. Las cifras de las medianas son tolerables, aunque algunas escalas de la porción baja del rango siguen siendo débiles. Las confiabilidades de consistencia interna de la mayoría de las escalas de contenido son adecuadas, lo cual también es comprensible si tomamos en cuenta que la homogeneidad de los reactivos fue uno de los criterios que se emplearon en la elaboración de estas escalas. La mayoría de la escalas de contenido también tiene una confiabilidad de test-retest adecuada, aunque no excepcionalmente alta. Podemos notar que los datos de confiabilidad son más favorables en el caso de las escalas de contenido que en el de las escalas clínicas, a pesar de que aquéllas son más cortas que éstas. El número medio de reactivos por escala es 24 en el caso de las escalas de contenido y 49 en el de las escalas clínicas.

Los datos de validez del MMPI-2 son demasiado extensos y complejos para permitirnos presentar un resumen útil en un breve espacio. Necesitamos consultar los manuales originales y la literatura de investigación para comprender una gran cantidad de información. Sin embargo, queremos señalar que los estudios analítico-factoriales sugieren que el MMPI-2 mide sólo cuatro dimensiones subyacentes. Las dos dimensiones mejor medidas son la actividad mental psicótica y las tendencias neuróticas (Butcher *et al*., 1989; Graham, 2006).

MMPI-2 RF (Forma reestructurada)

Muchos años de usar el MMPI-2, así como el MMPI original, ha llevado a algunos investigadores a desear cambios fundamentales en el instrumento. Dos características del MMPI provocaron una preocupación especial. Primero, el empirismo crudo del método de criterio meta para elaborar las escalas clínicas –la categoría más distintiva de las numerosas escalas del MMPI– deja inquietos a algunos usuarios. Más allá de mostrar una diferencia cruda entre un grupo clínico y otro no clínico, ¿las escalas tienen coherencia o un significado sustancial? Segundo, todas las escalas clínicas están llenas de una especie de factor general de desadaptación. ¿Se puede retirar este factor de las escalas clínicas y aislarse para hacer de éstas medidas más puras y únicas de sus respectivos objetivos?

Tratar con estas cuestiones ha llevado a la elaboración y publicación del MMPI-2 Forma reestructurada (**MMPI-2 RF**; Tellegen, Ben-Porath, McNulty, Arbisi, Graham, & Kaemmer, 2003; Ben-Porath & Tellegen, 2008a, 2008b; Tellegen & Ben-Porath, 2008). Por lo común, una nueva edición de una prueba implica nuevos reactivos, nuevas normas y nuevos programas de investigación, entre otras cosas, pero el MMPI-2 RF de ningún modo es una nueva edición en alguno de estos sentidos. No tiene reactivos nuevos ni se recolectaron datos nuevos para su elaboración. Se tomaron 338 reactivos de los 567 del MMPI-2, y las normas, análisis de reactivos y otras investigaciones provienen simplemente del proceso de los datos existentes del MMPI-2 de los 338 reactivos que

aparecen en nuevas configuraciones.

Lo realmente nuevo del MMPI-2 RF es el desarrollo de las escalas, en especial las escalas clínicas emblemáticas. Analizando nuevamente los reactivos y datos del MMPI-2, los autores del MMPI-2 RF esperaban alcanzar estas metas:

- Tener una puntuación separada de desadaptación general y tratar de minimizar su influencia en las otras escalas clínicas; ésta se convierte en la primera escala clínica: RCd, donde "d" representa la desmoralización.
- Purificar las otras escalas clínicas mediante varias estrategias psicométricas (análisis de reactivos, análisis factorial, etc.).
- Brindar otras escalas importantes más allá de las escalas clínicas reestructuradas.

Las escalas clínicas constituyen la contribución más distintiva del MMPI, aunque, como describimos antes, tiene una gran cantidad de otras escalas. Los fundamentos del MMPI-2 RF tienen que ver principalmente con las escalas clínicas; sin embargo, también ofrecen muchas otras escalas. En particular, el MMPI-2 RF preserva las nociones de escalas de validez y de contenido, aunque la elaboración exacta de éstas es un poco diferente en esta nueva versión. También se siguen usando los "reactivos críticos", pero con un nuevo nombre: "respuestas críticas", y se informan sólo cuando las escalas de las que derivan muestran elevación. En total, el MMPI-2 RF ofrece 50 escalas; el cuadro 13-6 presenta las principales categorías de estas escalas, sin enumerar todos los nombre de las escalas individuales, y ofrece breves descripciones de dichas categorías.

Cuadro 13-6. Bosquejo de las escalas del MMPI-2 RF

Escalas de validez
Ocho escalas de validez similares a las del MMPI-2 (véase cuadro 13-2)
Escalas de orden alto
Tres puntuaciones de resumen basadas en el análisis factorial de las escalas clínicas: disfunción emocional, disfunción del pensamiento, disfunción conductual
Escalas clínicas reestructuradas RC
Nueve puntuaciones: RCd es la escala general de "desmoralización"; otras ocho corresponden a las escalas tradicionales del MMPI (véase cuadro 13-3), sin las escalas 5 (masculinidad-feminidad) y 0 (introversión social)
Escalas de problemas específicos (SP), con cuatro subcategorías
Somática/Cognitiva: cinco puntuaciones de aspectos como dolor de cuello, problemas de memoria
Internalizante: nueve puntuaciones de aspectos como ideación suicida, ansiedad
Externalizante: cuatro puntuaciones de aspectos como agresión, abuso de sustancias
Interpersonal: cinco puntuaciones de aspectos como problemas familiares, timidez
Escalas de interés
Dos puntuaciones: estética-literaria, mecánica-física
Cinco escalas de psicopatología de la personalidad (PSY-5)
Cinco puntuaciones: agresividad, psicoticismo, falta de control, neuroticismo, introversión (recuerdan, en parte, a las escalas del NEO-PI; véase cuadro 12-12)

Recordemos nuestra descripción anterior de los métodos para elaborar medidas de personalidad, que incluyen el de criterio meta, el análisis factorial y el de contenido puro. Las escalas clínicas del MMPI original y el MMPI-2 se caracterizaron por el método de criterio meta, mientras que, en el caso del MMPI-2 RF, podemos escuchar el redoble de los tambores por el análisis factorial y el método de contenido, que se usaron en su elaboración.

Nuestro propósito principal aquí es presentar al lector las características generales de esta nueva encarnación del MMPI. El examen de todos los detalles técnicos de la forma reestructurada (el manual técnico es de 406 páginas, ¡la mayoría con cuadros de datos!) está más allá de nuestro propósito. Concluimos con algunas preguntas obvias. ¿El MMPI-2 RF es una prueba más seductora? ¿Reemplazará al MMPI-2 como el instrumento clínico más usado? ¿"Atrapará" a todos los usuarios? Algunos piensan que sí y otros, que no, pero sólo el tiempo lo dirá. Un número especial completo del *Journal of Personality Assessment* (Meyer, 2006) presentó los bandos opuestos. Una cosa parece cierta: le tomará mucho tiempo al MMPI-2 RF acumular la base de investigación que el MMPI-2 ha desarrollado con el paso de los años. El legendario récord del MMPI-2, con miles de estudios que tratan de cualquier tema y subgrupo imaginables, respalda muchos de los usos de esta prueba.

El Inventario Clínico Multiaxial de Millon (MCMI) y la familia Millon

Theodore Millon es el autor principal de una familia de pruebas objetivas de personalidad, cuyos miembros actuales se enumeran en el cuadro 13-7. Algunas son revisiones de otras pruebas, a veces con un cambio en el número de edición (p. ej., I, II), pero otras veces con un título por completo nuevo. Así, referirse "al Millon" es bastante confuso; sin embargo, el miembro más usado de la familia es el *Inventario Clínico Multiaxial de Millon* (**MCMI**, siglas en inglés), al que solemos referirnos cuando hablamos "del Millon", si bien hay aún cierta ambigüedad, ya que disponemos de tres ediciones (I, II y III). De acuerdo con Camara *et al*. (2000), el MCMI es, en la actualidad, el décimo instrumento más usado por psicólogos clínicos (el MMPI es el segundo y el WAIS el primero).

Cuadro 13-7. Familia de inventarios Millon

Inventario	Acrónimo	Fecha de publicación
Millon Clinical Multiaxial Inventory	MCMI	1976
Millon Clinical Multiaxial Inventory—II	MCMI-II	1987
Millon Clinical Multiaxial Inventory—III	MCMI-III	1994
Millon Adolescent Personality Inventory	MAPI	1982
Millon Adolescent Clinical Inventory	MACI	1993
Millon Behavioral Health Inventory	MBHI	1974

Millon Behavioral Medicine Diagnostic	MBMD	2001
Millon Index of Personality Styles-Revised	MIPS	2003
Millon Pre-Adolescent Clinical Inventory	M-PACI	2005
Millon College Counseling Inventory	MCCI	2006

El MCMI se ajusta a nuestro esquema de clasificación (cuadro 12-3) como inventario integral con orientación hacia la anormalidad, igual que el MMPI. De hecho, el MCMI ha surgido como el principal competidor del MMPI, pues, al igual que éste, está rodeado de toda una cultura. Para usarlo, se requiere el estudio cuidadoso de terminología especializada, procedimientos novedosos y supuestos subyacentes. En este capítulo sólo presentaremos los aspectos más destacados del MCMI.

El MCMI es un buen ejemplo de un método combinado en la elaboración de un inventario clínico. Hay tres hilos que corren a lo largo de su elaboración. Primero, empieza con una orientación hacia la teoría de la personalidad propuesta por Millon. Los reactivos originales se crearon para reflejar aspectos de esta teoría. Desde su primera publicación (Millon, 1969), la teoría ha sido reelaborada y reformulada en varias ocasiones; a veces se ha denominado teoría del aprendizaje biosocial, que postula tres polaridades básicas de la personalidad humana: placer-dolor, sí mismo-otros y activo-pasivo. También postula un continuo que va de la personalidad normal a la anormal, en el que distintos trastornos graves de personalidad son expresiones extremas de las polaridades. Millon (1981) identificó ocho patrones básicos de personalidad, y luego agregó tres variantes más graves de los patrones básicos. Las tres polaridades, los ocho patrones y sus variantes se manifiestan de varias maneras en las escalas del MCMI, aunque se han hecho cambios de una edición a otra. En la elaboración del MCMI también se emplearon valoraciones clínicas de los reactivos en términos de lo apropiado de las categorías de contenido. Por último, los reactivos se sometieron a un análisis estándar para mejorar la homogeneidad de las escalas.

El MCMI tiene tres características distintivas. Primero, y quizá lo más importante para el crecimiento de su popularidad en los años recientes, hace un intento explícito para ajustar sus puntuaciones al Manual Diagnóstico y Estadístico (DSM) de la *American Psychiatric Association*. Como señalamos antes, el DSM en sus varias ediciones ha ofrecido el principal sistema diagnóstico y terminológico de los trastornos mentales en el campo de la salud mental. Por ello, el MCMI, y cualquier otra prueba que se apegue al DSM, será atractivo para los usuarios.

Segundo, el MCMI es mucho más corto que el MMPI. Mientras que el MMPI-2 tiene 567 reactivos y su aplicación requiere cerca de una hora, el MCMI-III tiene sólo 175 reactivos de verdadero-falso y su aplicación requiere cerca de 25 min. El MCMI-III produce 26 puntuaciones, pero recientemente se han agregado 42 nuevas puntuaciones llamadas "Escalas de Facetas Grossman", que son una especie de subpuntuaciones dentro de otras puntuaciones. Contando éstas, tenemos la sorprendente cantidad de 70 puntuaciones derivadas de 175 reactivos de verdadero-falso en el MCMI-III. Tercero, el MCMI emplea puntuaciones de "tasa base", que toman en cuenta, al señalar áreas

problemáticas, la tasa base de los tipos de problemas psicológicos que hay en la población. Por ejemplo, el padecimiento X tiene una mayor prevalencia (una tasa base superior) que el padecimiento Y. Así, una escala para el padecimiento X marca más casos de problemas que la escala para el padecimiento Y. En contraste, en la mayoría de las pruebas, un puntuación alta o de problema se suele definir en un sentido estrictamente normativo; por ejemplo, una puntuación T de 65 o mayor se puede considerar problemática sin importar las diferentes tasas base en la población. Estas puntuaciones de tasas base son un desarrollo muy interesante.

El cuadro 13-8 resume las principales categorías de puntuaciones del MCMI-III y presenta ejemplos de algunas escalas dentro de cada categoría. Los que conocen el DSM reconocerán la compatibilidad de esta prueba con la terminología de dicho manual. Los Índices de modificación del MCMI-III son tipos de índices de validez.

Cuadro 13-8. Principales categorías de las puntuaciones del MCMI-III y ejemplos de sus escalas

Categoría	**Ejemplos de escalas**
Escalas de trastornos de la personalidad	
Moderado	Histriónica, Antisocial, Compulsiva
Grave	Limítrofe, Paranoide
Síndromes clínicos	
Moderado	Ansiedad, Bipolar: Maníaco, Dependencia del alcohol
Grave	Trastorno del pensamiento, Trastorno delirante
Índices de modificación	Revelación, Deseabilidad
Escalas de facetas Grossman	42 subpuntuaciones dentro de las escalas clínicas y de trastornos

¡Inténtalo!

Para tener acceso a informes muestra del MCMI-III, escribe “Pearson samplerpts” (forma abreviada de Pearson sample reports) en cualquier buscador de internet y sigue el vínculo. ¿Cómo se compara este informe con el del MMPI que aparece en la figura 13-1?

Las ventajas y desventajas del MCMI se han debatido con vehemencia. El lector debe consultar las reseñas del MCMI, que se citan más adelante, para leer un tratamiento minucioso de los temas pertinentes. Aquí, identificamos brevemente dos líneas argumentales principales. Primero, como señalamos antes, el MCMI es bastante breve en comparación con un inventario integral, pero aún así produce muchas puntuaciones, lo cual es muy atractivo para los usuarios. Sin embargo, la combinación de brevedad y multiplicidad de puntuaciones se consigue usando los mismos reactivos en diferentes escalas. En parte resultado de esta característica, algunas escalas tienen una correlación tan alta entre sí que es difícil justificar darles diferentes nombres. Es como tener eventos separados en una carrera de atletismo de 100 yardas y de 100 metros. ¿Por qué la

molestia? En sus varias ediciones, el MCMI ha luchado con esta situación. Segundo, como señalamos antes, el intento de ajustar las escalas del MCMI a los criterios del DSM ha sido muy atractivo para los clínicos. Sin embargo, se tiene que pagar un precio en dos aspectos. Primero, el DSM en sí mismo no es perfecto; en la medida en que la correspondencia de la prueba con el DSM sea exitosa, la prueba será imperfecta. Segundo, el DSM cambia; ahora se encuentra en su cuarta edición. El MCMI-II intentó ajustarse al DSM-III, y luego llegó el DSM-IV, por lo que el MCMI tuvo que ser revisado: apareció el MCMI-III. El MCMI se ha revisado de manera considerable en dos ocasiones en el tiempo relativamente corto en que ha estado disponible, lo que causa una tensión importante en la generalización de los primeros estudios a la edición actual. En la moderna sociedad occidental, tendemos a apreciar el desarrollo rápido de flamantes productos nuevos. En el caso de la elaboración de pruebas, a veces necesitamos reducir la velocidad para dejar que la investigación nos dé la información apropiada.

Todos los inventarios Millon que aparecen en el cuadro 13-7 tienen en común el método y el sello; tienen sus raíces en la teoría del aprendizaje biosocial de Millon con sus tres polaridades, aunque se sirven de esta teoría en grados variables. Todos combinan los métodos empírico y teórico en la elaboración de la prueba y también se caracterizan por los extensos informes narrativos y otros materiales complementarios. Con una excepción, fueron diseñados para usarse con grupos clínicos, tienen normas basadas en este tipo de grupos e incorporan algún tipo de tasa base en los informes. Sin embargo, más allá del MCMI que hemos descrito, tienen propósitos y grupos meta únicos.

El MMY de Buros contiene reseñas de particular utilidad del MCMI-III (Choca, 2001; Hess, 1998; Retzlaff, 1998; Widiger, 2001). Groth-Marnat (2009) ofrece una descripción minuciosa de la elaboración, investigación y esquema interpretativo del MCMI-III. El panorama general que presentan Millon y Davis (1996) es una revisión útil de los fundamentos de la prueba. En Millon (1997, 2008) y Strack (2008) también se puede encontrar información útil acerca de los inventarios Millon.

Symptom Checklist-90-R

Symptom Checklist-90-R (**SCL-90-R**; Derogatis, 1994) es uno de los instrumentos clásicos para una investigación rápida de autoinforme de diversos estados disfuncionales. Se trata de un buen instrumento para examinarlo, porque muchas otras medidas siguen un patrón similar y su número parece aumentar exponencialmente cada año. El grupo al que está dirigido el SCL-90-R son las personas de 13 años en adelante. Como lo sugiere su nombre, contiene 90 reactivos, cada uno de los cuales identifica un síntoma como sentimientos (de hecho, muchos reactivos empiezan con la palabra “siento”), patrones de pensamiento o conductas expresadas en una palabra o frase. Las respuestas se ubican en una escala de cinco puntos: 0 = Para nada, 1 = Un poco, 2 = Moderadamente, 3 = Bastante y 4 = Extremadamente. La aplicación típica toma cerca de 15 minutos. Los clientes usan “los últimos siete días incluyendo hoy” como el tiempo de referencia para sus respuestas.

El SCL-90-R es un descendiente directo del *Woodworth Personal Data Sheet*. Recordemos nuestra discusión del capítulo 1 acerca de este instrumento en la historia de las pruebas. Se creó como sustituto de una entrevista personal de exploración que se hizo a los reclutas del Ejército de EUA durante la Primera Guerra Mundial. La entrevista personal consumía bastante tiempo, y Woodworth sintió que se podía obtener mucha información crucial con un método sencillo de lápiz y papel. La escala de Woodworth fue muy usada, pero perdió popularidad a causa de la preocupación por la distorsión de las respuestas. El MMPI ganó popularidad, en parte, por su uso explícito de escalas de validez para detectar posibles distorsiones de las respuestas, como lo describimos antes. El SCL-90-R emplea los mismos fundamentos básicos que el Woodworth, es decir, obtiene información inicial con un formato de lápiz y papel que cubre una amplia gama de síntomas. Después se realiza un seguimiento según sea necesario; por ejemplo, se puede determinar (sin que el clínico invierta su tiempo) que una persona marcó "4 (Extremadamente)" en cinco reactivos relacionados con la ansiedad y quizá sólo en ellos. Sin duda, eso sugiere que el clínico debe empezar a investigar la ansiedad (y no las fobias ni otras posibles vías). Desde luego, hacer un diagnóstico va más allá de observar las puntuaciones de un inventario, pero éstas pueden indicar un punto de partida de una manera muy sencilla.

El SCL-90-R produce 12 puntuaciones. Nueve se denominan dimensiones primarias de síntomas y tres son índices globales. Los nombres, número de reactivos y descripciones breves de las *nueve dimensiones primarias de síntomas* son:

- Somatización (12): disfunciones corporales
- Obsesivo-compulsivo (10): pensamientos, acciones recurrentes
- Sensibilidad interpersonal (9): sentimientos de inadecuación, inferioridad
- Depresión (13): melancolía, muerte
- Ansiedad (10): tensión, pánico
- Hostilidad (6): ira, agresión, resentimiento
- Fóbico (7): temores injustificados
- Ideación paranoide (6): desconfianza excesiva
- Psicoticismo (10): retraimiento, aislamiento, alucinación

El lector suspicaz notará que el número de reactivos de estas escalas asciende a 83 en total. Los otros siete reactivos se denominan simplemente "reactivos adicionales" y cubren síntomas variados como problemas de la alimentación y el sueño, pero que no conforman una categoría coherente.

- Los 90 reactivos producen los siguientes *tres Índices globales*:
- Índice de gravedad global (GSI): suma de respuestas de los 90 reactivos.
- Índice de la aflicción de síntomas positivos (PSDI): suma de respuestas dividida entre el número de respuestas diferentes de cero, tomada como indicador de la "aflicción promedio".

- Total de síntomas positivos (PST): número de respuestas diferentes de cero, tomado como medida de la amplitud de los síntomas.

El manual señala que el Índice de gravedad global es el "mejor indicador único del nivel o profundidad actual del trastorno [y] debe usarse en la mayoría de casos en que una medida de resumen única se requiere" (Derogatis, 1994, pp. 12-13).

El SCL-90-R usa puntuaciones T normalizadas para sus normas. El manual recomienda que una puntuación T de 63 sirva como punto de corte para identificar un caso de "riesgo positivo". (Vuelve a leer el capítulo 3 para recordar las puntuaciones T. Observa que, de acuerdo con el cuadro 3-1, una puntuación T de 63 corresponde al percentil 90.) Más que tener un único conjunto de puntuaciones T, el manual proporciona normas separadas basadas en los siguientes grupos: psiquiátricos de consulta externa, psiquiátricos internados, no pacientes (adultos) y adolescentes no pacientes. Cada grupo se subdivide, a su vez, en normas de hombres y de mujeres, lo que da un total de ocho diferentes grupos normativos. Así, el clínico puede elegir la norma más apropiada para interpretar las puntuaciones del cliente. Incluso puede ser apropiado considerar los resultados con base en más de un grupo de estandarización. Por ejemplo, una puntuación natural de 1.11 en Somatización corresponde a una puntuación T de 63 para una mujer no paciente, de modo que está al borde de considerarse como un problema; esa misma puntuación natural corresponde a una puntuación T de 52 para un paciente psiquiátrico internado, de modo que está dentro del rango normal de su grupo. Todos los grupos de estandarización son de conveniencia (véase capítulo 3), por lo que es difícil juzgar en qué grado son representativos de una población más amplia, aunque el manual hace un esfuerzo razonable para describir los grupos.

Las puntuaciones del SCL-90-R parecen tener una confiabilidad razonable de consistencia interna, en general alrededor de .85. La estabilidad de test-retest varía de .68 a .90, con una mediana de .80 y la mayoría de los valores entre .75 y .85. Sin duda, hay razón suficiente para tener cuidado respecto de la estabilidad de algunas escalas.

Una cuestión crucial acerca de la validez del SCL-90-R es la independencia de los nueve grupos de síntomas. El manual presenta los argumentos de que las nueve escalas son razonablemente independientes en el sentido factorial. Otras fuentes sugieren que puede haber sólo una o dos dimensiones subyacentes de psicopatología, crítica basada en el análisis factorial que también se ha dirigido al MMPI.

¡Inténtalo!

Empleando la escala de respuesta del SCL-90-R, crea dos reactivos que pienses que podrían indicar un problema psicológico. Se muestra un reactivo. Agrega dos reactivos cortos más.

	Para nada	Un poco	Moderadamente	Bastante	Extremadamente
	0	1	2	3	4
Miedo a conocer personas	0	1	2	3	4

_____	0	1	2	3	4
_____	0	1	2	3	4

Ejemplos de pruebas de dominio específico

A continuación identificamos varios ejemplos de medidas de dominio específico, pero antes de leer sobre estos ejemplos, quizá el lector quiera revisar las características en común de estas medidas, que se resumen en el cuadro 12-6. Literalmente se usan miles de medidas de este tipo; las pruebas que tratamos aquí están entre las más usadas.

Inventario de Depresión de Beck (BDI) «347-348a

El *Inventario de Depresión de Beck*–Segunda Edición (**BDI-II**, siglas en inglés; Beck, Steer, & Brown, 1996) es un excelente ejemplo de una prueba de dominio específico. Camara *et al*. (1998, 2000) informan que entre las pruebas que se usan en la evaluación de la personalidad y la psicopatología, el BDI-II se ubica en el sexto puesto en el caso de los psicólogos clínicos y en el segundo (sólo detrás del MMPI) en el de los neuropsicólogos.

De acuerdo con el manual (Beck *et al*., 1996, p. 1), el BDI-II es un "instrumento de autoinforme para medir la gravedad de la depresión en adultos y adolescentes". Consta de sólo 21 reactivos y se puede contestar en 5 o 10 min. El propio manual de la prueba es de sólo 38 páginas. Cada reactivo se responde en una escala graduada de cuatro puntos (0-3), que va más o menos de "no es un problema" a "es un gran problema". Los troncos de los reactivos son, en esencia, muy sencillos: palabras únicas o frases describen un síntoma depresivo. El cuadro 13-9 muestra un reactivo similar a los del BDI-II.

Cuadro 13-9. Reactivo parecido a los del BDI-II

Marque la afirmación que describa mejor cómo se ha sentido en las dos últimas semanas.

Preocupación

0. No estoy especialmente preocupado.
1. Me preocupo bastante a menudo.
2. Me preocupo por casi todo.
3. Me preocupo tanto que me hago daño.

Al sumar las respuestas en la escala de 0 a 3 de los 21 reactivos se obtiene una puntuación natural dentro del rango 0-63. Para describir el grado de depresión, el manual (p. 11) sugiere estas categorías para los siguientes rangos de puntuaciones:

0-13 Mínimo
14-19 Leve
20-28 Moderado
29-53 Grave

Los puntos de corte de estas descripciones se obtuvieron de manera empírica para

distinguir en grado óptimo entre varios grupos evaluados clínicamente. La interpretación de las puntuaciones depende de estas clasificaciones; en el manual no aparecen normas tradicionales (p. ej., percentiles o puntuaciones estándar).

El manual presenta datos de confiabilidad y validez basados en una muestra de 500 pacientes externos diagnosticados clínicamente de acuerdo con los criterios del DSM en cuatro sitios, así como en una muestra de 120 estudiantes de una universidad canadiense. El manual del BDI-II informa coeficientes alpha de .92 en el caso de la muestra de pacientes externos y de .93 en el de la muestra universitaria. La confiabilidad de test-retest que se informa es de .93 para una submuestra de 26 casos del grupo de pacientes externos, con un intervalo entre las aplicaciones de una semana.

Respecto de la validez, el manual del BDI-II informa correlaciones con varias pruebas que respaldan la validez convergente y discriminante. Intenta mostrar, en el caso de la validez discriminante, que la prueba no es una medida primordialmente de ansiedad. El análisis factorial sugiere que los reactivos del BDI-II cubren dos dimensiones, una etiquetada como Somática-Afectiva y la otra como Cognitiva.

Una característica interesante del BDI-II es el hecho de que el manual presenta las curvas características de reactivo (CCR) por *cada respuesta* a cada reactivo. Estas CCR no forman parte formalmente de la interpretación de las puntuaciones, pero ayudan al usuario a comprender cómo cada reactivo e incluso cada respuesta funcionan dentro del contexto de la puntuación total. En Arbisi (2001), Farmer (2001) y Groth-Marnat (2009) se pueden consultar reseñas del BDI-II.

Inventario de Trastornos de la Conducta Alimentaria (EDI)

El *Inventario de Trastornos de la Conducta Alimentaria-3* (**EDI-3**, siglas en inglés; Garner, 2004) constituye otro excelente ejemplo de un instrumento de dominio específico con aplicación clínica. Como el BDI-II, el EDI-3 se centra en un área específica de problemas. A diferencia del BDI-II, el EDI-3, como veremos más adelante, ofrece varias puntuaciones y no una sola. La segunda edición de esta prueba, el EDI-2, ha sido la medida más usada en el campo de los trastornos de la conducta alimentaria.

El EDI-3 es un instrumento de autoinforme "que busca medir los rasgos psicológicos o grupos de síntomas relevantes para el desarrollo y mantenimiento de los trastornos de la conducta alimentaria" (Garner, 2004, p. 4). El manual de la prueba distingue entre escalas relevantes relacionadas directamente con los trastornos de la conducta alimentaria y escalas relacionadas con rasgos psicológicos más generales que predisponen a estos trastornos. Esta bifurcación es una característica importante del EDI-3.

La prueba consta de 91 reactivos; de manera extraña, un reactivo (71) no se califica. En el cuadro 13-10 aparecen los nombres de las escalas y el número de reactivos de cada una, así como las puntuaciones compuestas y los Indicadores del estilo de respuesta del EDI-3. Las puntuaciones compuestas y los indicadores del estilo de respuesta no existían en el EDI-2. Las puntuaciones compuestas son sumas de puntuaciones T de las escalas incluidas, que se convierten, a su vez, en otras puntuaciones T. Los indicadores del estilo

de respuesta son ejemplos clásicos de los "índices de validez" como los que encontramos antes en el MMPI-2. En el EDI-3 no hay superposición de reactivos entre las subescalas.

Cuadro 13-10. Estructura del Inventario de Trastornos de la Conducta Alimentaria-3

Puntuación compuesta	**Escala**	**Número de reactivos**
Riesgo de trastornos de la conducta alimentaria	Impulso por la delgadez	7
	Bulimia	8
	Insatisfacción corporal	10
Ineficacia	Baja autoestima	6
	Alienación personal	7
Problemas interpersonales	Inseguridad interpersonal	7
	Alienación interpersonal	7
Problemas afectivos	Déficits interoceptivos	9
	Disregulación emocional	8
Exceso de control	Perfeccionismo	6
	Ascetismo	7
Desadaptación psicológica general	8 escalas de las anteriores	(57)
	Miedo a la madurez	8
Indicadores del estilo de respuesta		
Inconsistencia	10 pares de reactivos	
Infrecuencia	10 reactivos	
Impresión negativa	Todos los reactivos (excepto el 71)	

Las escalas en la puntuación compuesta de Riesgo de trastornos de la conducta alimentaria se relacionan directamente con dichos trastornos, mientras que el resto de las escalas se relaciona con rasgos más generales. El manual hace referencia a éstas más adelante como escalas psicológicas, las cuales tienen varias puntuaciones compuestas que, al sumarse, dan como resultado la puntuación de la escala de Desadaptación psicológica general.

El cuadro 13-11 presenta un reactivo muestra del EDI correspondiente a la escala de Insatisfacción corporal, que formó parte del EDI-2 y continúa en el EDI-3. Las respuestas se hacen en una escala de seis puntos que va de "Siempre" a "Nunca"; en la mayoría de los reactivos, una respuesta en la dirección de alta frecuencia es "sintomática". En el caso de estos reactivos, la puntuación es: Siempre = 4, Habitualmente = 3, A menudo = 2, A veces = 1 y las otras dos respuestas (Rara vez y Nunca) = 0. En cerca de la tercera parte de los reactivos, una respuesta de baja frecuencia es sintomática; así, la puntuación tiene un peso inverso en estos reactivos. Las puntuaciones dentro de las escalas son resultado de las sumas de estos valores de las

respuestas, que son un poco diferentes respecto del EDI-2.

Cuadro 13-11. Sample Item from the Eating Disorder Inventory

45. I think my hips are too big. A U O S R N (A = Always U = Usually O = Often S = Sometimes R = Rarely N = Never)
Fuente: Reproducido con permiso especial de la editorial Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, de The Eating Disorder Inventory-3, por David M. Garner, Ph.D., Copyright © 1984, 1991, 2004 por PAR, Inc. Cualquier otra reproducción está prohibida sin la autorización previa de PAR, Inc.

El manual del EDI-3 proporciona las puntuaciones T y los rangos percentiles de los tres diferentes grupos: muestra clínica de adultos de EUA (N = 983), muestra clínica internacional (N = 662) y muestra clínica de adolescentes de EUA (N = 335). Estas muestras clínicas contenían casos de anorexia nerviosa–tipo restrictivo, anorexia nerviosa–tipo compulsivo/purgativo, bulimia nerviosa y trastorno de la conducta alimentaria no especificado. Estas cuatro categorías corresponden al DSM-IV. Todos los casos en los grupos de estandarización fueron de mujeres. Los dos grupos de EUA se obtuvieron de cinco estados, con la gran mayoría de casos provenientes de estados del medio oeste. La muestra internacional provenía de Australia, Canadá, Italia y Holanda.

Las muestras son una buena ilustración de los grupos de estandarización por conveniencia, pero los del EDI-3 son mucho más extensos que los del EDI-2, los cuales provenían en su totalidad de Toronto, Canadá. El manual hace referencia a muestras no clínicas de mujeres (también llamadas "controles"), pero ofrece poca información sobre la naturaleza de estos grupos. Por último se presentan las medias y desviaciones estándar de grupos clínicos y no clínicos de hombres, pero los cuadros normativos completos de estos grupos no aparecen. De hecho, estos datos se presentan en la sección del manual sobre la validez en vez de la correspondiente a las normas.

Las puntuaciones naturales de las escalas se grafican en un perfil que muestra en áreas sombreadas el rango de que va del percentil 25 al 67 en el caso de los grupos normativos clínicos y no clínicos. La figura 13-2 muestra un perfil de éstos. Podemos notar que las puntuaciones de este caso caen con claridad en el rango del grupo clínico, a excepción de la puntuación de Insatisfacción corporal, la cual cae en el rango del grupo no clínico. También podemos notar una elevación extrema en las escalas Bulimia y Disregulación emocional. Éste es un método inusual pero interesante para informar las puntuaciones de una prueba.

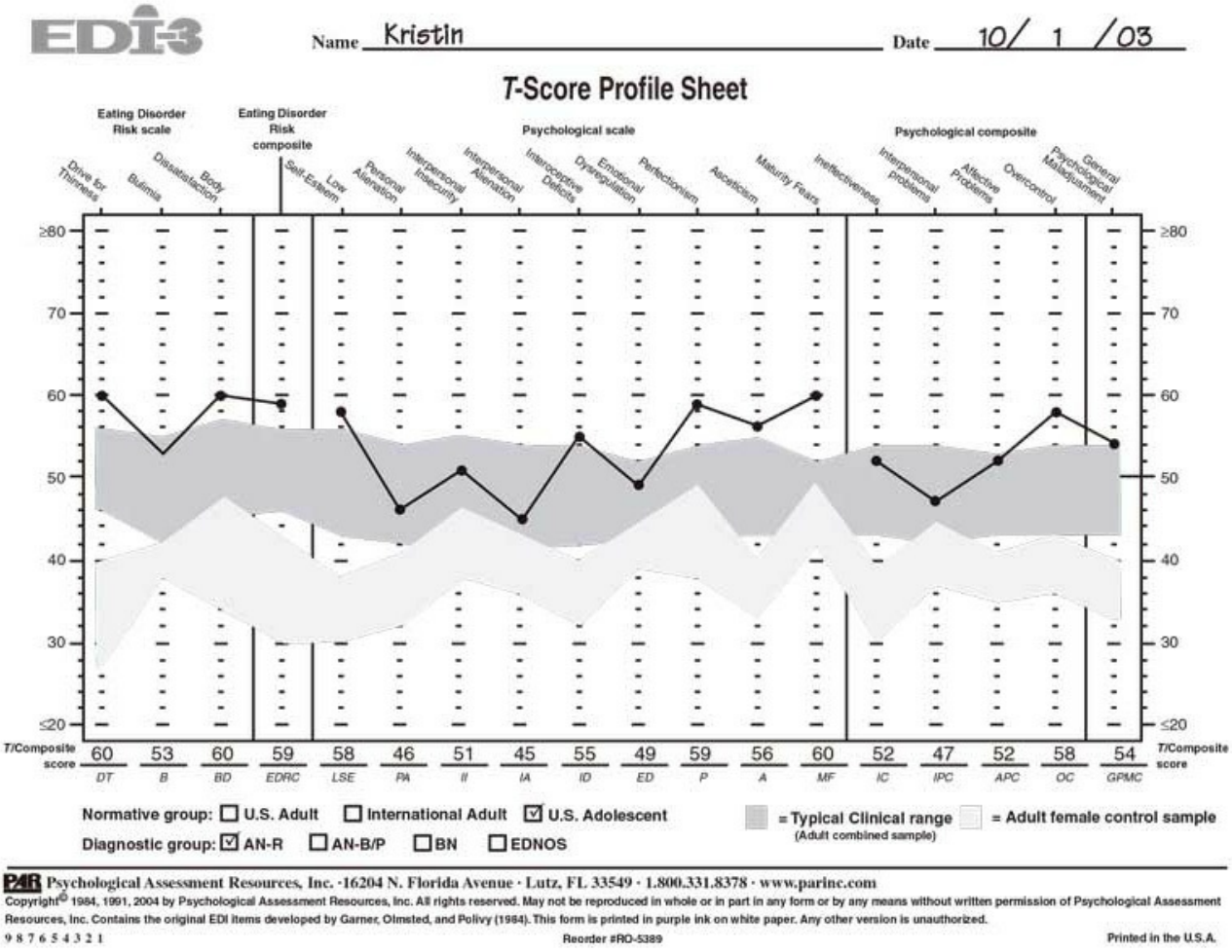


Figura 13-2. Perfil muestra del EDI-3.
Reproducido con permiso especial de la editorial Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, Florida 33549, del Eating Disorder Inventory-3, por David M. Garner, Ph.D., Copyright © 1984, 1991, 2004 por PAR, Inc.
Cualquier otra reproducción está prohibida sin la autorización previa de PAR, Inc.

¡Inténtalo!

Examina el perfil del EDI-3 que aparece en la figura 13-2. Compara el área sombreada en gris más oscuro con el área en gris más claro. ¿En qué escalas parece haber mayor separación entre los grupos de normalización clínicos y de control? ¿En qué escalas la separación es menor?

El manual del EDI-3 informa las confiabilidades de consistencia interna (alpha) de los tres grupos de estandarización clínicos que describimos antes. Además, informa las confiabilidades alpha y de test-retest de otras fuentes, cuyos coeficientes, en su mayoría, se ubican en el rango de .80 a .95, aunque ciertas escalas, en algunas muestras, tienen coeficientes alrededor de .65. La mayoría de las confiabilidades alpha es excelente; las de

test-retest también son altas, aunque sólo un estudio basado en 34 casos proporcionó estos datos. El manual también presenta bastante información sobre la validez; incluye correlaciones con muchas otras pruebas, un amplio trabajo analítico-factorial, en especial relacionado con la nueva estructura de la prueba, y estudios de grupos contrastados. Como suele ocurrir, es una tarea abrumadora revisar y dar sentido a cientos de correlaciones presentadas para sustentar la validez de la prueba.

El EDI-3 constituye un ejemplo clásico de métodos múltiples en la elaboración de una prueba. Primero, clínicos conocedores de los trastornos de la conducta alimentaria crearon un fondo de reactivos: método de contenido. Después, los reactivos se sometieron al método de criterio meta contrastando las respuestas a los reactivos de un grupo con trastornos de la conducta alimentaria y otro de no pacientes. Además, se determinaron las correlaciones reactivo-subescala para asegurar la homogeneidad de la subescala; esto es, en esencia, similar al método analítico-factorial. De hecho, posteriormente se llevó a cabo un análisis factorial para confirmar la relativa independencia de las subescalas y la coherencia de las puntuaciones compuestas.

Una Lista de control de síntomas (EDI-3 SC) complementa la forma principal de la prueba; indaga en temas como hacer dieta y ejercicio y la historia menstrual. Con la tercera edición también apareció el EDI-3 Forma de Referencia; se trata, en esencia, de una forma corta que consta de 25 reactivos de la prueba principal. Su objetivo es ser un dispositivo de exploración rápida para detectar casos que ameriten una evaluación más detallada con la versión completa del EDI-3.

Inventario de Ansiedad Rasgo-Estado

El último ejemplo de los instrumentos de dominio específico es el Inventario de Ansiedad Rasgo-Estado (**STAI**, siglas en inglés; Spielberger, 1983). Piotrowski y Lubin (1990) informan que el STAI es, sin duda, el instrumento más usados por los psicólogos de la salud en la evaluación de la ansiedad. El STAI también es una de las pruebas más citadas en la literatura especializada relacionada con todo tipo de pruebas (Plake & Impara, 1999). Como su nombre lo indica, el STAI intenta distinguir entre ansiedad estado (A-Estado) –un padecimiento temporal, transitorio– y ansiedad rasgo (A-Rasgo) –una característica más permanente, duradera.

El STAI usa 20 reactivos en un lado de la forma de la prueba para medir A-Estado y otros 20 en el otro lado para medir A-Rasgo. Los reactivos son muy similares, aunque el análisis de reactivos para la selección de éstos identificó cierta diferenciación en su naturaleza. El cuadro 13-12 presenta reactivos similares a los de las escalas A-Estado y A-Rasgo, así como los formatos de respuesta.

Cuadro 13-12. Reactivos parecidos a los del STAI

	Para nada	Un poco	Moderadamente	Mucho
Me siento tranquilo.	[1]	[2]	[3]	[4]

	Casi nunca	A veces	A menudo	Casi siempre
Me siento desdichado.	[1]	[2]	[3]	[4]

La diferencia crucial entre los dos conjuntos de reactivos es la dirección de las respuestas. En el caso de los de A-Estado, el examinado responde en términos de "qué tan bien se siente *justo ahora*, es decir, *en este momento*". En el caso de los reactivos de A-Rasgo, el examinado responde en términos de "cómo se siente *por lo general*"; el manual señala que se puede pedir al examinado que responda en términos de una situación particular, pero no "justo ahora", por ejemplo, cómo se siente justo antes de un examen. Las respuestas se califican del 1 al 4 y se suman para obtener una puntuación total de cada escala. En la mayoría de los reactivos, una respuesta de frecuencia alta indica ansiedad, pero algunas veces las de frecuencia baja indican ansiedad. Por lo tanto, en este último caso, el peso de la respuesta se invierte. La puntuación total de cada escala puede variar entre 20 (ansiedad baja) y 80 (ansiedad alta). Las puntuaciones promedio de estudiantes de bachillerato, universitarios y trabajadores adultos tienden a estar entre 35 y 40. El STAI no usa índices de validez; en su lugar, el manual simplemente sugiere que el examinador establezca *rapport* con el examinado y haga hincapié en la importancia de responder con honestidad. El manual también sugiere que si las circunstancias nos hacen esperar respuestas que no sean francas, se debe usar otra prueba que tenga índices de validez en vez del STAI.

El STAI está diseñado para usarse con estudiantes de bachillerato, universitarios y adultos. También existe una Escala de Ansiedad Rasgo-Estado para Niños (STAI-C; Spielberger, 1973), pero no la comentaremos aquí. Se presentan las normas –rangos percentiles y puntuaciones T– de varias muestras por conveniencia, entre las cuales se incluyen estudiantes de bachillerato, universitarios, reclutas del Ejército, adultos normales, pacientes neuropsiquiátricos, pacientes de medicina general y de cirugía, y presos. Las categorías de las normas del STAI ofrecen un excelente ejemplo de la necesidad de estar atentos para distinguir las normas basadas en muestras por conveniencia. Por ejemplo, las normas de los "estudiantes universitarios" se basan por completo en estudiantes de un curso introductorio de psicología en una institución, mientras que las de "adultos normales" se basan por completo en empleados de una agencia federal.

La elaboración del STAI es un caso interesante de cómo un instrumento evoluciona a lo largo de una serie de refinamientos basados primordialmente en procedimientos de análisis de reactivos. Cada etapa de este análisis lleva a ajustes en busca del propósito de la prueba, en este caso distinguir entre la ansiedad estado y rasgo. La confiabilidad de consistencia interna (alpha) de las escalas A-Estado y A-Rasgo en varias muestras tiene un promedio de .90 o mayor, lo cual no es sorpresivo, ya que la investigación en la elaboración de la prueba incluyó esfuerzos excepcionales para asegurar correlaciones reactivo-prueba altas. El STAI constituye un buen caso de estudio de cómo es posible obtener confiabilidad de consistencia interna alta por medio de un constructo bien definido y los procedimientos adecuados de análisis de reactivos.

La confiabilidad de test-retest del STAI constituye un caso inusual, pues se esperaría una confiabilidad alta en el caso de la escala de A-Rasgo. Sin embargo, si la A-Estado es, en realidad, un fenómeno transitorio, no se esperaría una confiabilidad alta de test-retest. En efecto, esto es lo que tiende a ocurrir. En varias muestras, esta confiabilidad en la escala A-Rasgo tiende a estar cerca de .75, mientras que en la escala A-Estado tiende a estar cerca de .30. Por lo común, este último nivel de confiabilidad sería un impresionante desastre psicométrico, pero, de acuerdo con la lógica de la medida de A-Estado, es comprensible e incluso deseable.

La validez del STAI se divide en cuatro categorías principales. Primero, hay correlaciones con otras pruebas, en especial las que pretenden medir ansiedad. Segundo, hay contrastes grupales, en especial de grupos en los que podría argumentarse que deberían tener patrones diferentes en las puntuaciones de A-Estado y A-Rasgo. Tercero, están los estudios del efecto de causas temporales de estrés, en especial para mostrar que las puntuaciones de A-Estado se elevan mientras que las de A-Rasgo se mantienen estables en tales situaciones. Por último, están los estudios analítico-factoriales diseñados para mostrar la separación de los reactivos de las dos escalas. No es sorprendente que algunos investigadores duden de la validez de separar la ansiedad-estado de la ansiedad-rasgo. De ahí que el manual del STAI dedique un espacio considerable para referirse a este tema. En Chaplin (1984), Dreger (1978), Groth-Marnat (2009) y Katin (1978) se pueden encontrar revisiones de esta prueba.

Escalas de valoración conductual

Ahora nos dedicaremos a un grupo de instrumentos clínicos que merecen estar en una categoría aparte. Las **escalas de valoración conductual** (EVC) se han vuelto muy populares para determinar padecimientos como trastornos de atención, hiperactividad, depresión y problemas emocionales de distinta índole. Tienen dos características esenciales. Primero, alguien diferente de la persona evaluada hace la valoración; por lo general, ese "alguien" es un maestro, uno de los padres u otro cuidador. De este modo, estas escalas son como las de adaptación conductual que discutimos en relación con la discapacidad intelectual en el capítulo 8. La única diferencia es que la escala de adaptación está dirigida a las habilidades funcionales como comer o habilidades sencillas de consumidor, mientras que las EVC están dirigidas a problemas como la hiperactividad.

La segunda característica esencial es que las EVC, como lo sugiere su nombre, enumeran conductas específicas. Esto también es similar a las escalas de adaptación conductual. La persona que contesta la forma indica la frecuencia con que se observa cada conducta. Los descriptores conductuales suelen ser cortos: de una a tres palabas. El cuadro 13-13 muestra descriptores típicos que pueden aparecer en una de estas escalas. Podemos notar que los reactivos intentan concentrarse en conductas específicas observables. En contraste, los inventarios de autoinforme, como el BDI o el EDI, a menudo se concentran en sentimientos o percepciones. Las valoraciones se hacen en escalas de 3 a 5 puntos, y por lo general varían de "Nunca" a "Siempre" o de "Totalmente falso" a Totalmente verdadero". La parte de la derecha del cuadro 13-13 muestra este esquema de valoración.

Cuadro 13-13. Ejemplos de reactivos típicos de las escalas de valoración conductual

	0	1	2	3
El niño...	Nunca	A veces	A menudo	Siempre
1. Golpea a otros niños	0	1	2	3
2. Está quieto	0	1	2	3
3. Hace trabajos descuidados	0	1	2	3
4. Llora	0	1	2	3
5. Le grita a otros	0	1	2	3
6. Termina tarde su trabajo	0	1	2	3
7. Hace ruidos extraños	0	1	2	3
8. Se retuerce cuando está sentado	0	1	2	3

Es práctico considerar dos amplios grupos de EVC. El primero incluye varios sistemas de puntuaciones múltiples, que intentan abordar varios padecimientos y, por lo común, producen una docena de puntuaciones o más. El segundo grupo incluye instrumentos que tienen una sola área como objetivo. Describamos brevemente ejemplos de cada

categoría.

Sistemas de puntuaciones múltiples

Existen tres escalas de valoración conductual de puntuaciones múltiples que se usan mucho (véase cuadro 13-14); pueden usarse otras, pero éstas son las que entran en acción con mayor frecuencia. Aquí bosquejamos las características en común de estas escalas, en vez de hacer una descripción detallada de cada una. En Ramsay, Reynolds y Kamphaus (2002) y Andrews, Saklofske y Janzen (2001), así como en los manuales de cada una de ellas (Achenbach &Rescorla, 2001; Conners, 2008; Reynolds & Kamphaus, 2004), puede encontrarse un tratamiento detallado.

Cuadro 13-14. Las tres escalas de valoración conductual de puntuaciones múltiples más usadas

Nombre	**Acrónimo**
Behavior Assessment System for Children	BASC, BASC-2
Child Behavior Checklist[a]	CBCL
Conners' Rating Scale, ahora Conners'3[b]	CRS o Conners-3

[a] Parte del Achenbach System of Empirically Based Assessment (ASEBA).
[b] Ahora también hay un Conners Comprehensive Behavior Rating Scale, así como otras escalas Conners.

Cada sistema de puntuaciones múltiples es, en realidad, un conjunto de varios instrumentos. Por lo común, los sistemas tienen escalas separadas que los padres, maestros y el niño contestan. La forma para el niño es, en realidad, un instrumento de autoinforme como los que describimos antes, es decir, el niño se describe a sí mismo, mientras que a padres y maestros se les pide valorar la conducta del niño. Además, algunos de estos sistemas tienen formas largas y cortas, así como niveles para niños más pequeños o mayores. Así, decir que se empleó "el Conners" o "el BASC" puede ser ambiguo, pues podría significar una forma larga o corta para padres o maestros o un autoinforme del niño. Necesitamos tener cuidado ante estas referencias.

La característica más importante del sistema de calificaciones múltiples es su intento de abarcar problemas de muchas áreas. Algunas formas producen varias docenas de puntuaciones. El cuadro 13-15 enumera áreas que con frecuencia abarcan estos sistemas. Desde luego, cada uno tiene su conjunto único de puntuaciones, pero las áreas que aparecen en el cuadro son su esencia.

Cuadro 13-15. Ejemplos de áreas que abarcan las escalas de valoración conductual de puntuaciones múltiples

Agresión	Ira
Ansiedad	Depresión
Hiperactividad	Inatención
Oposición	Retraimiento

Escalas de área única

Existen numerosas escalas de valoración conductual de área única; sus reactivos y escalas de respuesta son del mismo tipo que los del cuadro 13-13. Sin embargo, como lo sugiere su nombre "área única", se concentran en problemas de una sola área, por ejemplo, inatención. Así, tienden a ser mucho más cortas que las de puntuaciones múltiples. Mientras que uno de estos instrumentos puede tener cerca de 100 reactivos, una escala de área única puede tener sólo 20. Igual que dichos sistemas, estas escalas, por lo común, son contestadas por maestros, padres u otros cuidadores.

Las escalas de valoración conductual, tanto las de sistemas de puntuaciones múltiples como las de área única, se usan mucho en la actualidad en contextos educativos. Ayudan a identificar niños con problemas especiales, así como a cuantificar el grado del problema, y sirven como medida de seguimiento para mostrar el progreso del tratamiento del problema.

Resumen de puntos clave 13-3

Principales categorías de técnicas de evaluación conductual
Observación directa/en ambientes naturales
Observación conductual análoga
Pruebas situacionales
Role-playing
Entrevista conductual
Automonitoreo y autoinforme
Evaluación cognitivo-conductual
Medidas fisiológicas

Evaluación conductual

La evaluación conductual no es una prueba específica ni un método particular como los reactivos de opción múltiple o las técnicas proyectivas; más bien representa un método general o, incluso, una filosofía para obtener información sobre características humanas importantes. El "argumento de la evaluación conductual" dice así: si queremos información sobre depresión, fobias o ansiedad, ¿por qué hacer una serie de preguntas que ofrecen sólo información indirecta? ¿Por qué no observar directamente estas características? Por ejemplo, observar casos de conducta deprimida, fóbica o ansiosa en un contexto social. Los defensores de la evaluación conductual la contrastan con la *evaluación tradicional*, que en este contexto significa evaluación con pruebas como el MMPI, el NEO PI o el Rorschach. La teoría de la evaluación conductual tiene tres principios fundamentales. Primero, piensa en las conductas observables más que en rasgos subyacentes. Segundo, para medir un tipo particular de conducta hay que estar tan cerca de ella como sea posible y, además, observar los antecedentes inmediatos y las consecuencias de la conducta, es decir, lo que sucede antes y después de ella. Tercero, en la medida de lo posible, tratar de relacionar la medición con lo que se desea hacer con la información. Por ejemplo, si medimos algo para remediar un problema, se debe ligar la medición con el método para remediar el problema. Los dos primeros principios son los más importantes desde el punto de vista teórico, mientras que el tercero es más importante desde el punto de vista práctico, en especial, es apto para usos clínicos de la medición. La evaluación conductual tiene un gran encanto intuitivo, que quizá sea engañoso.

Las raíces de la evaluación conductual se pueden rastrear en dos corrientes de pensamiento dentro de la psicología. Primero, fue una reacción contra las teorías psicodinámicas, que hacían hincapié en el origen remoto de los problemas; por ejemplo, el clínico con orientación psicodinámica buscaría los orígenes de una fobia en la infancia de la persona. Pero los clínicos reaccionaron contra esta teoría pensando que quizá se podría eliminar una fobia sin ahondar en la infancia. Al mismo tiempo, aumentó la impaciencia en relación con los rasgos tan generalizados que medían algunas pruebas de lápiz y papel y algunas técnicas proyectivas (p. ej., Mischel, 1968). Segundo, la evaluación conductual se desarrolló junto con la terapia conductual; esto fue parte de la revolución conductista en la psicología clínica, que consistió en llevar las teorías del aprendizaje de Pavlov y Skinner al campo de los problemas clínicos: extinguir una fobia mediante los principios de Pavlov, o reemplazar una respuesta por otra mediante los principios de Skinner. Emplear estos conceptos y técnicas en la terapia demandó medidas de distinta clase, es decir, el clínico necesitaba medidas de las reacciones fisiológicas y conteos de conductas específicas en vez de reflexiones sobre una mancha de tinta. Se puede consultar Goldfried y Kent (1972) y Goldfried (1976) para conocer el sello distintivo de los primeros escritos de la evaluación conductual, y Groth-Marnat (2009) y Haynes, O'Brien y Kaholokula (2011) para conocer los cambios en esta perspectiva que

han ocurrido con el paso de los años.

La evaluación conductual busca ser sumamente específica de acuerdo con las circunstancias particulares de cada caso. Por lo tanto, no encontramos pruebas muy usadas en esta área, sino que tenemos categorías de técnicas que deben adaptarse a cada caso. En esta sección, presentamos las principales categorías (véase Resumen de puntos clave) e ilustramos brevemente una aplicación de cada una. Algunas categorías son muy distintas de las demás, mientras que los límites entre otras son difusos.

Resumen de puntos clave 13-3

Principales categorías de técnicas de evaluación conductual
Observación directa/en ambientes naturales
Observación conductual análoga
Pruebas situacionales
Role-playing
Entrevista conductual
Automonitoreo y autoinforme
Evaluación cognitivo-conductual
Medidas fisiológicas

Observación directa o naturalista

La observación directa o en ambientes naturales puede considerarse la técnica emblemática de la evaluación conductual. Consiste en observar la conducta en el ambiente en que ocurre de manera natural. Consideremos el caso de un niño que manifiesta conductas agresivas en la escuela; podríamos tener un observador en el salón de clases anotando la frecuencia, tipo, duración e intensidad de las conductas agresivas del niño. El psicólogo escolar podría usar estas observaciones como línea base para indicar el progreso en el tratamiento del caso; también puede diseñar un programa de reforzamientos, tanto negativos como positivos, para reducir las conductas meta. También podemos considerar el caso de una persona petrificada entre una multitud de personas. Podríamos observar a la persona mientras se acerca a una zona con mucha gente y registrar las conductas exactas en esa situación. Después de aplicar un tratamiento, podríamos observar otra vez a esta persona en un contexto social para ver si ha ocurrido un cambio conductual.

Observación conductual análoga

La observación directa, aunque atractiva, tiene evidentes inconvenientes. Por lo general, es muy poco práctica y cara. Además, podemos desconocer cuándo ciertos tipos de conductas ocurrirán; por ejemplo, una persona puede sufrir una crisis de pánico cada dos semanas, lo cual puede causar estragos en su vida, pero no sabemos cuándo se

presentará la siguiente crisis. Una pareja de casados puede pelearse una vez al mes, lo cual pone en peligro su matrimonio, pero no sabemos cuándo observar la siguiente pelea. La **observación conductual análoga** intenta simular la observación directa; Haynes (2001) la definió de esta manera: "La observación conductual análoga implica la observación de clientes (p. ej., niños, adultos, familias, parejas) en un ambiente diseñado para aumentar la probabilidad de que el evaluador pueda observar conductas e interacciones clínicamente importantes" (p. 3). Incluimos en esta categoría las **pruebas situacionales** y el ***role-playing***. Una prueba situacional implica colocar a la persona en una situación que se aproxima a otra en la que nos gustaría hacer una predicción; por ejemplo, supongamos que queremos predecir qué tan bien funcionará una persona en un puesto que requiere colaborar con otras personas. Desde luego, podríamos hacer una predicción con base, digamos, en las puntuaciones del NEO PI. En la evaluación conductual se colocaría a la persona en una reunión simulada de un comité y se observaría su conducta. Incluso podríamos tener un cómplice en la reunión que, de manera deliberada, trate de provocar a la persona o que altere el curso de la reunión. Entonces, las demás personas que forman parte de ella u observadores externos pueden valorar el funcionamiento de la persona en esa situación. La **discusión grupal sin líder** es una aplicación de este método. En este caso, el grupo tiene un tema de discusión, pero nadie se encarga de conducirla; entonces se observa cómo reacciona una persona en esta situación. Podríamos tener un interés particular en determinar si la persona que evaluamos impone cierto liderazgo.

En el *role-playing*, pedimos a una persona asumir un papel; por ejemplo, en un programa de entrenamiento en asertividad, podríamos pedir a la persona ser firme y adoptar una actitud frente a cierto tema en un contexto grupal. Podemos observar cómo se desenvuelve la persona en esta tarea y, después, analizar la experiencia con ella y repetirla para ver si se ha hecho algún progreso. Podemos notar que no estamos profundizando en el inconsciente de la persona; sólo tratamos de evaluar directamente la conducta y cambiarla.

Entrevista conductual

La entrevista puede ser un elemento sorpresivo en una lista de técnicas de evaluación conductual. De hecho, no aparecía en las primeras listas, pero ahora es una parte normal. Sin embargo, la entrevista conductual es muy diferente de lo que se suele considerar una entrevista; en ella, de acuerdo con los principios de la evaluación conductual que describimos antes, el entrevistador se concentra en detalles de conductas específicas. Consideremos el caso de una persona que tiene dificultades para controlar su carácter; el entrevistador puede preguntar por ejemplos específicos: así que usted explotó con su compañero de cuarto, ¿cuándo sucedió esto?, ¿había alguien más presente?, ¿qué hizo su compañero?, ¿qué estaba haciendo justo antes de explotar?, ¿cuánto tiempo duró este arranque?, y así sucesivamente. Se podría hacer una serie de preguntas detalladas como estas a una persona que sufre de crisis de pánico: ¿dónde estaba?, ¿qué estaba haciendo

justo antes de la crisis?, ¿a qué hora fue? En esencia, el entrevistador intenta aproximarse a la observación directa.

Automonitoreo y autoinforme

Otro modo de aproximarnos a la observación directa es pedir a la persona que haga registros cuidadosos de su conducta. Cone (1999) definió el **automonitoreo** como "el acto de observar y registrar de manera sistemática aspectos de la propia conducta y los eventos internos y externos que, se piense, estén relacionados funcionalmente con la conducta" (p. 411). Los clínicos usan técnicas de automonitoreo de manera habitual (Korotitsch & Nelson-Gray, 1999). Los registros pueden ser conteos (cuántos cigarros fumó en un día) o categorías de actividades. Un ejemplo interesante del uso de categorías implica el **método de muestro de experiencias** (Csikzentmihalyi & Larson, 1987). Una persona tiene una lista de categorías de actividades. Durante el día, en períodos aleatorios, escucha un timbre y debe registrar lo que hace en ese momento. Esto se puede usar, por ejemplo, con un estudiante para determinar cuánto tiempo dedica efectivamente al estudio. Si se trata de contar los cigarros que se fuman en un día, la meta es reducir la cantidad de cigarros. La persona también podría registrar las circunstancias en que enciende un cigarro, lo cual podría ayudar a determinar qué suele provocar el aumento en la conducta de fumar. Al estudiar la actividad de estudiar, la meta es aumentarla; al anotar otras actividades, el estudiante sabrá qué estuvo haciendo en vez de estudiar: ver televisión o pasar el rato con sus amigos podrían ser los culpables.

La mayoría de las listas de técnicas de evaluación conductual incluye el autoinforme, que se acerca peligrosamente a los métodos de evaluación tradicionales, como los que revisamos en el capítulo 12 y antes en este capítulo. De hecho, demuestra que los métodos de evaluación conductuales y tradicionales se ubican a lo largo de un continuo antes que ser por completo diferentes. En el contexto de la evaluación conductual, el autoinforme requiere informes detallados de las conductas y sus antecedentes y circunstancias.

Evaluación cognitivo-conductual

En la *evaluación cognitivo-conductual* se tratan los pensamientos de la persona además de sus conductas. La noción básica es que el problema se debe a lo que una persona piensa de una situación; si hay un problema, quizá se deba a los pensamientos más que a la conducta externa. Miedos, baja autoestima, depresión: todos tienen que ver con los pensamientos. Por ello, el remedio es cambiarlos, así que queremos que los pensamientos salgan a la superficie.

Al igual que con otras técnicas de evaluación conductual, el truco es ser sumamente específicos. Un método es la **técnica de hablar en voz alta**: la persona se imagina en una situación fóbica y verbaliza todos sus pensamientos. También se puede llevar a cabo

un tipo de entrevista conductual para evocar estos pensamientos. Esto puede sonar muy parecido a la entrevista psicoanalítica tradicional, pero la principal diferencia es la orientación; en el contexto cognitivo-conductual, la orientación es en el aquí y ahora, y no en el pasado lejano, y hacia el cambio en los patrones de pensamiento, más o menos inmediatamente, y no hacia la comprensión de los procesos inconscientes. En Glass y Arnkoff (1997) se puede encontrar una discusión detallada de este método.

Medidas fisiológicas

Las medidas fisiológicas constituyen uno de los mejores ejemplos de la evaluación conductual; son el regocijo del conductista de hueso colorado. De acuerdo con el conductista, el miedo es sólo un haz de reacciones fisiológicas, igual que la ansiedad ¡y el amor! De ahí que, si queremos medir el miedo o la ansiedad, medimos las reacciones fisiológicas. Consideremos, por ejemplo, el miedo a hablar en público. El miedo incluye un ritmo cardiaco elevado, aumento en la conducción electrodérmica, pupilas contraídas, etc. Para tratar con el miedo, necesitamos establecer una tasa base de estas medidas, usar técnicas para modificar las reacciones fisiológicas y, por último, medirlas otra vez. Si tenemos éxito, el ritmo cardiaco y la conducción disminuirán, las pupilas ya no estarán contraídas, etc. Por lo tanto, el miedo habrá desaparecido. Podríamos usar técnicas de retroalimentación biológica, junto con técnicas cognitivo-conductuales de pensar en voz alta para lograr los cambios fisiológicos. En este ejemplo, podemos notar la interacción entre evaluación y tratamiento.

Conclusiones sobre los métodos de evaluación conductual

Muchos de los primeros escritos sobre evaluación conductual fueron optimistas e incluso ingenuos en exceso. Básicamente se decía: mide la conducta de manera directa y no tendrás que preocuparte por cuestiones de confiabilidad y validez. Además, daba la impresión de que este tipo de evaluación era sencillo. Sin embargo, pronto fue evidente que estas medidas a menudo *no* eran sencillas y que la confiabilidad y validez *eran* muy importantes. La evaluación conductual puede requerir demasiado tiempo y ser muy incómoda, además de que no es válida y confiable de manera automática.

Ahora, los defensores de la evaluación conductual se dan cuenta de que sus técnicas deben cumplir con los mismos requerimientos que las medidas tradicionales, los cuales se relacionan con confiabilidad, validez, nomas, así como eficacia y conveniencia. El desarrollo actual de esta área muestra un saludable interés por estos temas, como se refleja en los números especiales recientes de *Psychological Assessment* acerca del automonitoreo (Cone, 1999) y la observación conductual análoga (Haynes, 2001). En Haynes y Kaholokula (2008), Haynes, O'Brien y Kaholokula (2011) y en varios capítulos de McKay (2008) se puede encontrar un tratamiento completo de las técnicas de evaluación conductual.

Resumen de puntos clave 13-4

Tendencias en la elaboración y uso de instrumentos clínicos
Influencia del DSM y el CIE que está por venir
Énfasis en la planeación del tratamiento y la evaluación de seguimiento
Uso de instrumentos más breves
Crecimiento del número de instrumentos
Mayor uso de la aplicación en línea y de los informes interpretativos

Tendencias en la elaboración y uso de instrumentos clínicos

Nosotros sugerimos cinco tendencias principales en la elaboración y aplicación de las clases de instrumentos clínicos que describimos en este capítulo (véase Resumen de puntos clave).

Algunas de estas tendencias son las mismas que observamos en las pruebas de personalidad del capítulo anterior o, al menos, muy parecidas. Otras son muy diferentes.

La primera tendencia se relaciona con el predominio del DSM. En la actualidad, casi todos los instrumentos clínicos intentan ligar sus resultados con las categorías del DSM; de hecho, algunos las usan como punto de partida. Esta tendencia es tan evidente y abrumadora que es fácil pasarla por alto. ¡Pensemos en la lucha en torno del DSM-5! Quizá más importante, pensemos en el cambio radical que se presenta cuando los psicólogos (así como otros profesionales de la salud) cambian al sistema del CIE.

Segundo, aunque estos instrumentos clínicos están orientados primordialmente hacia el diagnóstico (en especial, de las categorías del DSM), el énfasis en la planeación del tratamiento y la evaluación del seguimiento están aumentando. Es decir, los resultados no sólo me dicen qué está mal (diagnóstico), sino también qué debo hacer (tratamiento). En el curso del tratamiento, también debo aplicar otra vez el instrumento para ver si hay progresos.

Tercero, ha surgido una tendencia al uso de instrumentos más breves. Para estar seguros, los instrumentos más extensos, en especial el MMPI-2, aún se usan mucho; sin embargo, dos fuerzas parecen alentar el uso de instrumentos más breves. La primera es la el manejo cuidadoso y su énfasis en la eficiencia del tiempo (entiéndase ahorro de dinero). La segunda, que mencionamos en el punto anterior, es el deseo de hacer una evaluación de seguimiento. No es difícil aplicar (ni contestar) una prueba como el BDI-II varias veces en pocos meses, pero ¿si se tratara del MMPI-2?

Cuarto, como se observó en las pruebas de rasgos normales de personalidad, somos testigos de una verdadera explosión en la publicación de instrumentos clínicos nuevos (o revisados). Esto es evidente sobre todo en el caso de los instrumentos breves. Aunque no se trata de un juego de niños elaborar y publicar una prueba de 20 reactivos enfocados en una área de problemas, tampoco es una tarea colosal. En cambio, elaborar un MMPI-2 o un MCMI requiere un esfuerzo y recursos sobrehumanos. El crecimiento del número de instrumentos nuevos hace hincapié en la necesidad de profesionales que conozcan los principios fundamentales aplicables a la evaluación de pruebas, es decir, los principios relacionados con confiabilidad, validez y normas.

Por último, la computarización del mundo continúa. Al igual que las pruebas de otras áreas, vemos que el uso de la aplicación (y calificación) en línea aumenta, así como los informes computarizados, en especial los interpretativos. Las versiones adaptadas por computadora de estas pruebas, sobre todo de las más extensas, parecen estar a la vuelta

de la esquina.

Resumen

1. En este capítulo se examinan los instrumentos clínicos y los métodos cuya meta primordial es identificar trastornos psicológicos. Estos instrumentos y métodos son similares en ciertos aspectos y diferentes en otros con respecto de los inventarios descritos en el capítulo anterior. Las semejanzas incluyen la naturaleza de los reactivos y formatos de respuesta, las subdivisiones prácticas en pruebas integrales y de dominio específico, las estrategias de elaboración y la preocupación por la dirección y el falseamiento de respuestas. Las diferencias giran en torno de la orientación general de su uso, los escenarios de aplicación y el interés en el diagnóstico, la planeación del tratamiento y el seguimiento.
2. El uso de algún tipo de entrevista clínica es casi universal. La entrevista tradicional no estructurada está llena de deficiencias técnicas, mientras que las entrevistas clínicas estructuradas intentan mitigar estos problemas. El ejemplo más conocido es el *Structured Clinical Interview for DSM-IV Axis I Disorder* (SCID-I), que intenta usar una secuencia cuidadosa de preguntas estandarizadas que llevan a las categorías diagnósticas del DSM. Por cierto, mucho de lo que se dijo acerca de la entrevista clínica se aplica igual de bien a la entrevista de trabajo.
3. El MMPI-2 es el instrumento de autoinforme más usado en la evaluación clínica. Su origen, estructura, puntuaciones, uso en investigación e incluso su vocabulario son legendarios en el campo de la psicología clínica. Constituye el principal ejemplo del método de criterio meta para la elaboración de pruebas, aunque sus puntuaciones actuales van más allá de dicho método. También ilustra una aplicación muy clara de los "índices de validez" y los "tipos de código" en el proceso del informe. Recientemente, también está disponible el MMPI-2 Forma Reestructurada.
4. Varios inventarios intentan proporcionar un examen integral de los trastornos. En primer lugar se encuentran los inventarios de la "familia" Millon, de los cuales el MCMI-III es el más usado. Intenta hacer una articulación cuidadosa con el DSM-IV. El SCL-90-R es uno de los inventarios integrales más breves; produce nueve puntuaciones primarias y tres índices globales.
5. Examinamos tres instrumentos de dominio específico, cada uno de un área específica de los problemas psicológicos. El Inventario de Depresión de Beck (BDI-II), el índice más usado para valorar los síntomas de depresión, es un instrumento sorprendentemente breve de sólo 21 reactivos. El Inventario de Trastornos de la Conducta Alimentaria (EDI-III) intenta ofrecer información de trastornos como anorexia y bulimia, pero también sobre padecimientos psicológicos que pueden exacerbar o conducir a estos trastornos. El Inventario de Ansiedad Rasgo-Estado (STAI) intenta distinguir entre la ansiedad temporal y la más duradera.
6. Las escalas de valoración conductual son notablemente diferentes de los otros instrumentos descritos en este capítulo. Una persona diferente de la persona evaluada se encarga de contestar estas escalas, por ejemplo, uno de los padres, un cuidador o un

maestro. Hacen hincapié en conductas muy específicas.

7. Las técnicas de evaluación conductual incluyen varios métodos que intentan examinar la conducta de una manera más directa que los inventarios de autoinforme. Identificamos los siguientes ejemplos de estas técnicas: observación directa o en ambientes naturales, observación conductual análoga, entrevista conductual, automonitoreo, evaluación cognitivo-conductual y medidas fisiológicas. Aunque estas técnicas aspiran a ser medidas más directas que los inventarios de autoinforme, también tienen que demostrar su confiabilidad y validez, así como resolver cuestiones de carácter práctico.

8. Identificamos cinco tendencias de los instrumentos clínicos que describimos en este capítulo. El DSM ha ejercido una gran influencia en la elaboración y estructura de estos instrumentos. Se hace un notable énfasis en la planeación del tratamiento y la evaluación de seguimiento, en parte como resultado de las demandas de la atención administrada. Quizá también como resultado de estas demandas, vemos un aumento del uso de instrumentos más breves. Hay un gran aumento en el número de instrumentos clínicos que se elaboran en la actualidad, sobre todo de pruebas de dominio específico. Por último, debido a la mayor aplicación de la tecnología de cómputo, aumenta el uso de la aplicación en línea de las pruebas y la disponibilidad de informes interpretativos extensos.

Palabras clave

automonitoreo
BDI
CIE
discusión grupal sin líder
DSM
EDI
entrevista clínica estructurada
escala de valoración conductual (EVC)
MCMI
método de muestreo de experiencias
MMPI
MMPI-2 RF
observación conductual análoga
prueba situacional
role-playing
SCID
SCL-90-R
STAI
técnica de hablar en voz alta
tipo de código

Ejercicios

1. A continuación aparece un conjunto de puntuaciones T del MMPI-2. ¿Cuál es el *tipo de código de dos puntos* de esta persona?
Escala 1 2 3 4 5 6 7 8 9 0
Puntuación T: 45 65 50 75 52 48 63 59 48 52

2. Observa la muestra de reactivos del cuadro 13-13, es decir, reactivos que podrían aparecer en una escala de valoración conductual. Agrega cinco reactivos a la lista. Asegúrate de que los reactivos cubran conductas específicas, en especial las que puedan ser problemáticas.

3. Para consultar informes muestra del MMPI-2 y el MCMI-III, escribe "Pearson samplerpts" (observa que es una forma abreviada de "sample reports") en cualquier buscador de internet. Debes encontrar una gran cantidad de informes. Examina algunos de cada inventario. ¿Qué concluirías acerca de la naturaleza de estos informes?

4. Observa el reactivo del cuadro 13-9. ¿Qué otros reactivos con el mismo formato esperarías encontrar en una lista de síntomas de depresión? Haz una lista de tres de estos reactivos. Luego, si tienes acceso al cuadernillo del BDI-II, compara tus reactivos con los que aparecen realmente en la prueba.

5. Aquí hay un ejemplo de cómo se aplica la evaluación conductual. Identifica un tema que consideres que induce ansiedad leve, como hablar en público o presentar un examen final. Imagínate en esa situación y registra tus pensamientos exactos. Enumera los eventos inmediatos que preceden la situación. Trata de ser tan específico como sea posible.

6. Usa el sitio de internet de ETS Test Collection (http://www.ets.org/test_link/find_tests/). Introduce como palabra clave un trastorno psicológico, como *paranoia*, *anxiety* o *depression* (paranoia, ansiedad o depresión). ¿Cuántas entradas obtuviste? ¿Puedes decir algo sobre la calidad de las pruebas/inventarios a partir de lo que comprendiste?

CAPÍTULO 14
Técnicas proyectivas

Objetivos

1. Identificar las principales características de las técnicas proyectivas.
2. Describir la hipótesis proyectiva.
3. Identificar los principales usos de las técnicas proyectivas y cuáles son las que se emplean con mayor frecuencia.
4. Bosquejar las principales características del Rorschach, incluyendo el Sistema integral de Exner.
5. Bosquejar las principales características del Test de Apercepción Temática.
6. Bosquejar las principales características de las técnicas de frases incompletas y de dibujo de la figura humana.
7. Discutir los factores que afectan el uso futuro de las técnicas proyectivas.

Las técnicas proyectivas constituyen uno de los temas más fascinantes no sólo en el campo de las pruebas psicológicas, sino en toda la psicología. Se encuentran entre los símbolos más fácilmente reconocidos de la psicología en la sociedad contemporánea. ¿Quién no ha encontrado referencias a una mancha de tinta en películas, novelas o dibujos animados? Estas técnicas también se encuentran entre los temas más controvertidos de la psicometría. Para algunos, son denostables por absurdas, seudocientíficas y deberían mandarse a los deshechos de la psicología junto con la frenología. Para otros, son una rica fuente de comprensión y, en comparación con ellas, las respuestas Sí-No de los inventarios objetivos de personalidad son triviales, incluso degradantes. En Lilienfeld, Wood y Garb (2000) y en dos secciones especiales sobre el Rorschach de *Psychological Assessment* (Meyer, 1999, 2001) se puede encontrar buenos ejemplos de la controversia alrededor de las técnicas proyectivas. Aunque las secciones especiales se concentran en el Rorschach, muchos argumentos, en pro y en

contra, se pueden generalizar a otras técnicas proyectivas. En este capítulo, exploramos esta intrigante categorías de pruebas: exponemos sus fundamentos, describimos sus usos y presentamos ejemplos de los métodos que se emplean con mayor frecuencia.

Características generales de las técnicas proyectivas y la hipótesis proyectiva

Las técnicas proyectivas tienen dos características clave. Primero, los reactivos de las pruebas suelen ser **estímulos ambiguos** en cierto grado, pues no es claro de inmediato qué significan, lo cual contrasta con los reactivos de las pruebas objetivas de personalidad (p. ej., A menudo me siento triste) en los que el significado es razonablemente claro (aunque algunos pueden argumentar que la expresión "a menudo" está abierta a la interpretación). La segunda característica clave es que usan un formato de respuesta abierta, que también se conoce como de respuesta libre, lo cual también contrasta con las pruebas objetivas de personalidad, que emplean un formato de respuesta cerrada. Como hemos señalado en varios momentos a lo largo de este libro, el uso del formato de respuesta libre crea un reto especial para la calificación de la prueba.

El fundamento que subyace en las técnicas proyectivas se denomina a menudo **hipótesis proyectiva**: si el estímulo es ambiguo, la respuesta estará determinada por la dinámica de la personalidad de la persona. Hay poco en la naturaleza de los estímulos de las pruebas proyectivas que dicte qué sería una respuesta razonable. Entonces, ¿cómo puede el examinado formular una respuesta? De acuerdo con la hipótesis proyectiva, la respuesta se formula en términos de los deseos, fantasías, inclinaciones, temores y motivaciones de la persona. Así, se piensa que las pruebas proyectivas son el medio ideal para descubrir las características profundas, quizá inconscientes, de la personalidad. Es decir, las pruebas proyectivas pueden investigar a mayor profundidad, mientras que las objetivas sólo tocan las características superficiales de la personalidad; al menos así dice la hipótesis.

Aunque no es parte de la hipótesis proyectiva, el método psicoanalítico para explorar la personalidad a menudo es aliado de esta hipótesis. Muchos defensores de estas técnicas provienen de la tradición psicoanalítica; sin embargo, es posible ser partidario de la hipótesis proyectiva también desde otras perspectivas. Por ejemplo, algunas versiones del método Gestalt, que hacen hincapié en la interacción de la personalidad con la percepción, pueden simpatizar con la hipótesis proyectiva.

Usos de las técnicas proyectivas

Existen *dos usos principales* de las técnicas proyectivas. Primero, se usan en la evaluación de casos individuales en psicología clínica, y escolar y en consejería psicológica. Segundo, se usan en la *investigación*. Consideremos en primer lugar el uso aplicado; de acuerdo con encuestas a psicólogos sobre los usos de pruebas, las técnicas proyectivas ocupan los primeros lugares con una regularidad notable. Consideremos estos hallazgos; en una encuesta a psicólogos que trabajan con adolescentes, Archer, Maruish, Imhof y Piotrowski (1991) encontraron que siete de las 10 pruebas más usadas eran técnicas proyectivas. En otra encuesta a psicólogos escolares, Kennedy, Faust, Willis y Piotrowski (1994) encontraron que seis de las 10 pruebas más usadas eran técnicas proyectivas. De acuerdo con Watkins, Campbell, Nieberding y Hallmark (1995), entre los psicólogos clínicos, cinco de las siete pruebas que se usan con mayor frecuencia son técnicas proyectivas. Estos patrones se han encontrado de manera consistente durante mucho tiempo (Lubin, Larsen, & Matarazzo, 1984) y en diversos contextos (Lubin, Larsen, Matarazzo, & Seever, 1985). Incluso ante las predicciones de psicólogos sobre la disminución del uso de estas técnicas (Piotrowski & Keller, 1984), las encuestas más recientes muestran que siguen empleándose mucho (Camara, Nathan, & Puente, 2000).

En este punto, será útil identificar con exactitud qué pruebas proyectivas se encuentran en los primeros lugares; por lo regular, son las siguientes ocho "pruebas": Prueba Rorschach de Manchas de Tinta, Test de Apercepción Temática (TAT), Test de Apercepción Infantil (CAT), pruebas de frases incompletas, dibujos de la figura humana, Test Gestáltico Visomotor de Bender (o simplemente Bender), Prueba Árbol-Casa-Persona (HTP) y Dibujo Cinético de la Familia (KFD) (cuadro 14-1). De hecho, prácticamente ninguna otra técnica proyectiva se ubica en los primeros lugares en las encuestas de preferencia, aunque existen muchas otras.

Cuadro 14-1. Las ocho técnicas proyectivas más usadas

Prueba Rorschach de Manchas de Tinta	Pruebas de Frases Incompletas
Test de Apercepción Temática (TAT)	Dibujo de la Figura Humana
Test de Apercepción Infantil (CAT)	Prueba Casa-Árbol-Persona (HTP)
Test Gestáltico Visomotor de Bender	Prueba de Dibujo Cinético de la Familia (KFD)

Sin tratar de negar el gran uso de estas técnicas, hay varias características peculiares que, quizá, han llevado a exagerar su posición en las encuestas sobre el uso de pruebas. En efecto, considerar estas características brinda una excelente introducción al campo de las técnicas proyectivas, por lo que dedicaremos un poco de tiempo a explorarlas. Primero, los psicólogos a menudo las usan sólo de una manera muy informal. (Ahondaremos en este punto más adelante.) De hecho, pueden no calificarlas en un sentido formal y usar sólo parte de los estímulos materiales. Algunos materiales de estas pruebas pueden servir para **romper el hielo**, casi como parte de una conversación en la

entrevista clínica. Esto apenas coincide con el concepto de prueba psicológica aplicado a otros tipos de pruebas. La mayoría de las encuestas sobre el uso de pruebas no hace distinciones respecto de estos aspectos; sin embargo, algunas sí las hacen y sus resultados son reveladores. Por ejemplo, Kennedy *et al*. (1994) pidieron a los encuestados indicar las razones por las que usan una prueba. En un número significativo de casos, los encuestados dijeron que usan una técnica proyectiva para "romper el hielo"; muchos también dijeron que no usaban procedimientos de calificación "estandarizados" con estas pruebas. En el caso de algunas, la mayoría de los encuestados informó el uso de un sistema de calificación "personalizado" y algunos dijeron no usar un sistema de calificación. En contraste, pocos informaron usar las medidas objetivas para romper el hielo o recurrir a un método de calificación distinto al estandarizado. De ahí que, cuando se les pregunta sólo si usan una prueba particular, los encuestados quizá respondan "Sí" aun cuando usen la prueba de manera informal. En un giro interesante en el catálogo psicométrico usual de los usos de pruebas, Holaday, Smith y Sherry (2000) señalaron que "los usuarios también informaron usar [técnicas de frases incompletas] para obtener citas textuales que pudieran apoyar el diagnóstico en los informes psicológicos..." (p. 380).

Segundo, la mayoría de las técnicas proyectivas tiene varios sistemas de calificación. Exploraremos este punto con mayor detalle cuando nos ocupemos de estas técnicas más adelante en este capítulo. Ya que el sistema de calificación es en realidad parte de la prueba, cada uno debe representarse por separado en la encuesta. Quien usa el sistema Klopfer para calificar el Rorschach no usa la misma prueba que quien usa el sistema Exner; no obstante, la manera en que se lleva a cabo la mayor parte de las encuestas sobre el uso de pruebas toma en cuenta a todos los que usen el Rorschach sin importar el sistema de calificación que utilicen, lo cual contribuye a que esta prueba aparezca entre las más usadas. Los ejemplos más extremos de este fenómeno son las frases incompletas y el dibujo de la figura humana, que no son pruebas específicas, pues, en realidad, existen docenas de pruebas de frases incompletas diferentes. En muchas otras encuestas se agrupan como una categoría genérica: frases incompletas. Puede ser que ninguna de ellas se use tanto; sin embargo, cuando se agrupan en una categoría, ésta ocupa uno de los primeros lugares en las encuestas. Del mismo modo, existen muchas pruebas de dibujo de la figura humana y las encuestas tienden a agruparlas en una sola categoría, lo que la ubica como una de las más usadas.

Por último, señalamos que las técnicas proyectivas se usan, por lo común, en la evaluación de variables de personalidad. Sin embargo, en algunos casos se usan para propósitos muy diferentes, pero incluso en este caso, contribuyen a la posición en las encuestas de las pruebas proyectivas en comparación con otras pruebas. El mejor ejemplo de esta dificultad es el *Test Gestáltico Visomotor de Bender*, al que se suele hacer referencia simplemente como el Bender. Algunas encuestas lo clasifican explícitamente como prueba proyectiva (p. ej., Piotrowski & Keller, 1984; Watkins *et al*. 1988), porque, se supone, a menudo se usa para evaluar la personalidad. Sin embargo, en ciertas circunstancias, el Bender sirve primordialmente como examen neuropsicológico

para detectar disfunción cerebral. Este uso bifurcado se ilustra en un informe de Camara *et al*. (1998) en el que el Bender aparece entre los primeros lugares tanto en la evaluación de la personalidad como en la neuropsicológica. De modo similar, señalamos que la forma más popular de la prueba de dibujo de la figura humana fue diseñada originalmente para medir inteligencia, con un sistema de calificación específico, aunque en la actualidad se utilice principalmente como medida proyectiva de la personalidad. Así, el multifacético uso de algunas de estas pruebas también contribuye a que las técnicas proyectivas aparezcan entre las pruebas más usadas.

Además de su uso en el trabajo clínico aplicado, estas técnicas también se usan mucho en la *investigación*, dividida en dos categorías principales. Primero, hay una gran cantidad de investigación sobre las características psicométricas de las medidas proyectivas; en este caso se examina la confiabilidad y validez de las técnicas empleando una amplia variedad de grupos. Segundo, estas técnicas se usan, a menudo, como variable criterio; en este caso, se supone que la técnica posee confiabilidad y validez aceptables y, luego, se usa para definir las variables del campo de la personalidad o del funcionamiento intelectual.

¡Inténtalo!

Para comprender la variedad de pruebas de frases incompletas que están disponibles en la actualidad, escribe las palabras "sentence completion" en el sitio de internet de ETS Test Collection (http://www.ets.org/test_link/about). Observa la cantidad de entradas que resultan.

Indicadores del uso de las técnicas proyectivas

En el trabajo aplicado, existen circunstancias que pueden llevar al psicólogo a preferir una prueba proyectiva en lugar de una objetiva o, al menos, a incluir la prueba proyectiva en la batería que se aplica al cliente. Primero, la mayoría de las técnicas proyectivas no requieren leer a diferencia de los inventarios objetivos de personalidad (aunque en algunos se permite leer los reactivos al examinado). Así, si un examinado no puede leer o lee con mucha lentitud, una prueba proyectiva puede ser deseable. Segundo, aunque las pruebas proyectivas son susceptibles de ser falseadas (positiva o negativamente), es más difícil hacerlo en éstas que en los inventarios objetivos de personalidad. De esta manera, si se sospecha que el examinado tiene una fuerte motivación para falsear sus respuestas, puede ser preferible una prueba proyectiva. Tercero, muchas técnicas proyectivas permiten formular un rango excepcionalmente amplio de hipótesis acerca de la dinámica de la personalidad; así, si el psicólogo tiene una base inicial muy reducida para juzgar las dificultades de un cliente, puede ser beneficiosa una prueba proyectiva.

Resumen de puntos clave 14-1

Tres métodos generales para calificar pruebas proyectivas
Formal
Informal
Holístico/Basado en impresiones

Aplicación y calificación de técnicas proyectivas: advertencia

Cuando un psicólogo dice que usó el WISC-IV, es seguro suponer que la prueba se aplicó de acuerdo con los procedimientos estandarizados, que se aplicó la prueba entera y que se calificó de acuerdo con los criterios especificados en el manual del WISC. Si un psicólogo interpreta un perfil del MMPI, es seguro suponer que la prueba se calificó de la manera usual; de hecho, la prueba quizá se calificó mediante un programa de cómputo diseñado por la editorial de la prueba. En contraste, cuando escuchamos que se aplicó el Rorschach o el TAT, podemos suponer poco acerca de lo que eso significa. Ya que las instrucciones de aplicación de las técnicas proyectivas suelen ser muy sencillas, es probable que se hayan empleado instrucciones razonablemente parecidas a las que indica la prueba. Sin embargo, en algunos casos, no se usan todos los materiales de la prueba; por ejemplo, de las 20 láminas del TAT o de las 10 del Rorschach, puede haberse usado sólo algunas. La calificación presenta un panorama mucho más variado.

El análisis de la literatura de investigación y las descripciones de la práctica clínica revelan tres métodos generales para calificar las técnicas proyectivas (véase Resumen de puntos clave), que, quizá, forman parte de un continuo en la práctica real. El primer método implica la *calificación formal* de acuerdo con reglas establecidas; podemos llamarlo método cuantitativo o psicométrico. Se trata del mismo método que se usa para calificar el WISC, y produce puntuaciones específicas que pueden estar relacionadas con normas y están sujetas a estudios ordinarios de confiabilidad y validez. En el otro extremo se encuentra el uso *informal*, que no implica calificaciones ni conclusiones definidas.

Los estímulos proyectivos se usan de una manera muy semejante a los elementos de una entrevista. En este método, presentar algunas láminas del Rorschach equivale más o menos a preguntar "¿Cómo le va el día de hoy?" Con este uso se busca formular hipótesis a las que se dará un seguimiento más específico en el subsiguiente trabajo. De ahí que este uso a menudo se denomine generación de hipótesis; por ejemplo, el resultado de este uso informal puede ser simplemente sugerir la necesidad de una medida específica de depresión o de discutir sobre las relaciones familiares. El tercer método para calificar una técnica proyectiva implica llegar a alguna conclusión, por ejemplo, un diagnóstico basado en la *impresión general* que causan las respuestas del examinado en lugar del análisis de puntuaciones específicas. Por ejemplo, el clínico aplica el Rorschach o el dibujo de la figura humana de un modo estandarizado, pero no se utiliza el sistema formal de calificación; sin embargo, con base en la impresión holística causada por las respuestas, el clínico concluye que el examinado es esquizofrénico. Este método a

menudo se denomina calificación holística o basado en impresiones.

Nos referiremos a estos tres métodos de calificación cuando examinemos las técnicas proyectivas en las siguientes secciones. Debemos estar atentos a estos distintos métodos al leer artículos en revistas especializadas o informes clínicos que emplean técnicas proyectivas.

Prueba Rorschach de Manchas de Tinta

La Prueba Rorschach de Manchas de Tinta, también conocida como método Rorschach de manchas de tinta o técnica Rorschach de manchas de tinta, es sin duda la técnica proyectiva más usada. Independientemente de sus características específicas, ilustra muchos problemas a los que se enfrenta cualquier técnica proyectiva. Por estas dos razones, le dedicaremos más tiempo que a las otras técnicas que presentamos en este capítulo.

Materiales

Varias técnicas utilizan manchas de tinta como estímulos. Sin duda, la más famosa y usada es la que identificamos con Hermann Rorschach, psiquiatra suizo que experimentó con un conjunto de manchas de tinta a principios del siglo XX. Rorschach murió a la edad de 38 años, poco después de publicar su primer y único trabajo con las manchas de tinta. La perspicaz obra de Rorschach estaba, sin duda, en una etapa preliminar cuando él murió. Su conjunto de manchas de tinta fue la base de la mayor parte del subsiguiente trabajo con esta técnica, por lo que nuestra presentación se concentra en las manchas de tinta de Rorschach.

¡Inténtalo!

Casi todos los novatos de la psicometría escriben mal "Rorschach". Aunque no coincide con la pronunciación del nombre (Ror-shaj), podrías pensar en segmentarlo de esta manera: Rors-ch-a-ch. Escribirlo bien es signo (aunque muy menor) de madurez profesional.

Encuestas recientes ubican al Rorschach entre las pruebas usadas con mayor frecuencia (véase, p. ej., Camara *et al*., 2000; Frauenhafer *et al*. 1998). Craig y Horowitz (1990) pidieron a directores de centros de prácticas clínicas identificar pruebas sobre las que se deba dar una formación especial al estudiante de posgrado en psicología clínica. El Rorschach ocupó el primer lugar en esta encuesta.

El Rorschach consta de 10 manchas simétricas. (En realidad, algunas tienen asimetrías muy ligeras.) La figura 14-1 muestra una mancha similar a las primeras manchas de Rorschach. Cada mancha aparece en una lámina rígida de aproximadamente 14 × 21.5 cm, casi el tamaño de este libro, pero no tan gordo. Las láminas están numeradas del I al X en la esquina superior derecha de la parte posterior; esta numeración corresponde al orden estándar de presentación. La ubicación de los números de las láminas permite presentarlas al examinado en una orientación estándar. Una reproducción muy pequeña de la firma de Hermann Rorschach adorna la parte posterior de cada lámina en algunas ediciones. Los números romanos tienen especial importancia, porque la literatura sobre el

Rorschach está plagada de referencias a las respuestas típicas y atípicas a ciertas láminas que se identifican mediante estos números. Por ejemplo, un autor puede decir "la respuesta del cliente de 'dos pájaros volando' a la lámina III es muy inusual". El psicólogo experimentado en el uso de Rorschach puede relacionar una afirmación de este tipo.

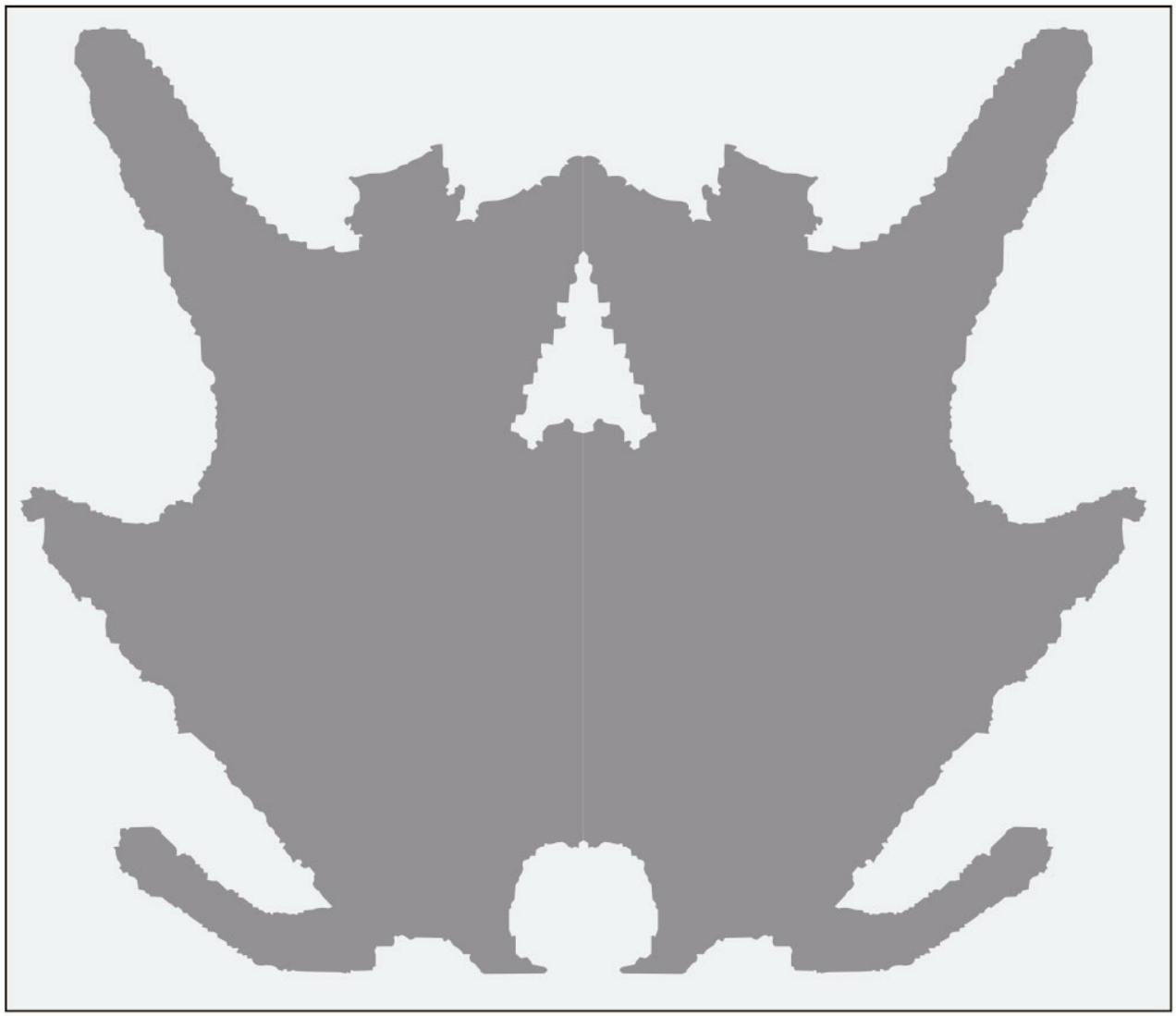


Figura 14-1. Mancha acromática de tinta similar a las del Rorschach.

La mayoría de las falsificaciones de las manchas de tinta del Rorschach muestra una mancha totalmente negra con un fondo blanco, quizá porque ésta es una manera muy fácil de reproducirlas. En realidad, ninguna mancha es completamente negra; algunas no tienen negro en absoluto. Cinco láminas (I, IV, V, VI y VII) son por completo acromáticas, pues contienen varios grados de gris y algunas partes negras. Dos láminas (II y III) son negras y grises en su mayor parte, pero tienen algunas manchas rojas. Las

últimas tres láminas son por completo cromáticas. Dos de ellas (VIII y IX) son combinaciones pasteles apagadas de rosa, verde y anaranjado. La última (X) tiene rosa, azul, amarillo y verde en abundancia.

¡Inténtalo!

En realidad, no lo intentes. Al leer acerca de las diversas técnicas proyectivas que presentamos en este capítulo, puedes tener el casi irresistible deseo de probar estos estímulos con tus amigos y conocidos para interpretar sus personalidades con base en sus respuestas. ¡No lo hagas! Se necesita una formación avanzada para usar estas técnicas. Como estudiante de pruebas psicológicas, tienes una responsabilidad especial. Un estudiante de biología o historia puede "jugar" con estas técnicas de manera inofensiva, porque ninguna persona razonable tomaría su interpretación con seriedad; sin embargo, si alguien sabe que estudias pruebas psicológicas, puede creer que estás calificado para hacer interpretaciones. Pero no lo estás, no hasta que hayas estudiado estos procedimientos más de lo que harás en este capítulo. Como en otros capítulos, tenemos varios ejercicios ¡Inténtalo! a lo largo de todo el texto, así como ejercicios al final del capítulo; sin embargo, en ninguno de ellos se te piden respuestas de otras personas a estímulos proyectivos ni analizar la personalidad de otras personas.

Aplicación y calificación «366a

El libro de Hermann Rorschach de 1921, *Psychodiagnostik*, no presentó un conjunto estándar de instrucciones para aplicar o calificar las 10 manchas de tinta. Después de la muerte de Rorschach en 1922, durante varias décadas, distintos psicólogos estadounidenses desarrollaron *sistemas*, es decir, instrucciones para aplicar y calificar las manchas de tinta de Rorschach. Hubo cinco sistemas, cada uno identificado en la literatura con el nombre del principal arquitecto del sistema. Beck (1937), Klopfer (1937; Klopfer & Kelley, 1942), Hertz (1943, 1948), Rapaport (Rapaport, Gill &, Schafer, 1946) y Piotrowski (1937, 1957). Podemos notar que las primeras referencias de todos estos sistemas surgieron en el periodo relativamente corto que va de 1937 a 1946. También existía la Técnica de Manchas de Tinta de Holtzman (Holtzman, 1961); podemos notar que "Rorschach" no aparece en el nombre, lo cual fue absolutamente deliberado por parte de Holtzman. Quería usar manchas de tinta pero de una manera por completo diferente respecto de otros sistemas del Rorschach; por ejemplo, él permitía sólo una respuesta por cada mancha de tinta. No obstante, la técnica de Holtzman a menudo se agrupa con los otros sistemas del Rorschach. Aiken (1999) ofrece un bosquejo histórico breve pero interesante de los orígenes de estos sistemas, cada uno de los cuales tuvo cierta actualidad, pero ninguno se convirtió en el estándar. De hecho, tener todos estos sistemas que, en parte, se superponían y, en parte, entraban en conflicto creó confusión. Recordemos nuestro tratamiento de la *Escala Wechsler de Inteligencia para Niños* del capítulo 8. Imaginemos que hay varios conjuntos diferentes de instrucciones para aplicar esta prueba, diferentes modos para calificar las respuestas y diferentes perfiles de puntuaciones. El resultado sería un caos; esa es la situación que ha

prevalecido por muchos años en lo que respecta al Rorschach.

Otro psicólogo estadounidense, John Exner, Jr. (Exner, 2003), creó lo que se denomina **Sistema integral** para aplicar y calificar las manchas de tinta del Rorschach.[1] Exner intentó incorporar en su Sistema integral las mejores características, las más justificables y, en apariencia, las más provechosas de los cinco sistemas. El de Exner se ha convertido en el sistema estándar en años recientes. Hiller *et al*. (1999) se refieren a la "adopción casi universal del Sistema integral del Rorschach de Exner" (p. 292). Hilsenroth y Handler (1995) informaron que el sistema de Exner para usar el Rorschach se enseñó a 75% de los estudiantes de posgrado. Este sistema será el que describamos; sin embargo, el lector debe tener en mente que, cuando busque en la literatura, puede encontrar alguno de los otros sistemas. Recordemos también que, en la práctica, un psicólogo puede usar un sistema "personalizado" o puede no utilizar ninguno.

Los procedimientos para aplicar el Rorschach en el marco del Sistema integral de Exner son sencillos. La aplicación se divide en dos fases: la fase de respuesta (también conocida como fase de asociación o libre asociación) y la fase de indagación. En la **fase de respuesta**, el examinador da la lámina al examinado y le pregunta "¿Qué podría ser esto?" Si el examinado busca alguna orientación sobre cómo responder o lo que está permitido (p. ej., voltear la lámina), el examinador es no directivo usando respuestas como "Eso depende de usted" o "Lo que a usted le parezca". Si el examinado es demasiado breve, por ejemplo, da una respuesta de una sola palabra (p. ej., "una mariposa"), el examinador lo debe animar a dar respuestas completas. Por ejemplo, el examinador puede decir: "La mayoría de las personas ve más de una cosa". Cada una de las 10 manchas se presenta con estas sencillas instrucciones. El examinador se sienta junto al examinado y le da la lámina al examinado.

Mientras el examinado responde sobre una mancha, el examinador registra lo que dice tomando nota del tono emocional. Exner hace hincapié en que las respuestas deben registrarse de manera textual. También se anota el tiempo de respuesta. Estos registros serán importantes no sólo para la subsiguiente calificación de respuestas, sino también para la fase de indagación.

En la **fase de indagación** se presentan al examinado otra vez las 10 manchas. Ahora, el examinador usa sus notas de la fase de respuesta y le pide al examinado explicar y elaborar más sus respuestas de la fase anterior. El examinador puede decir: "Muéstreme dónde vio [ESPACIO]". Desde luego, el examinador también registra de manera textual las respuestas del examinado de esta fase. Exner hace hincapié en que el propósito de la fase de indagación es ayudar en la codificación de las respuestas obtenidas en la primera fase y no evocar respuestas por completo nuevas.

El registro de respuestas se denomina **protocolo**. Este término se usa en la literatura de la mayoría de técnicas proyectivas, aunque a veces también se aplica fuera de este campo, por ejemplo, el protocolo de respuestas del MMPI, pero por lo general corresponde a los métodos proyectivos.

La aplicación del Rorschach es sumamente sencilla; sin embargo, la calificación no lo es en absoluto. De hecho, la calificación, cuyo nombre oficial es **codificación**, es muy

elaborada y detallada. Aquí esbozamos sólo cuestiones generales del esquema de codificación del Sistema integral. Consideremos las siguientes respuestas a una mancha imaginaria:

Fase de respuesta: Aquí veo algo que parece un animal muerto, algo así como si hubiera sido atropellado. Y supongo que esto que se ve rojo es sangre, así que tuvo que haber pasado hace poco tiempo. En esta parte de aquí, parece que algo está fluyendo, tal vez se trata de un *puddle* saliéndose del camino.

Fase de indagación: Aquí (señala la parte central de la mancha) está el animal muerto y aquí, la sangre manando a chorros. Aquí (tocando la parte periférica de la mancha) hay agua fluyendo hacia afuera del camino, probablemente en un arroyo al lado del camino.

¿Qué hace el psicólogo con estas respuestas? El punto más importante que se debe comprender es que el psicólogo no llega de inmediato a conclusiones basadas en estas respuestas. La imagen popular de interpretar las respuestas del Rorschach puede ser un método que se aplica por sí solo rociado de repentinos destellos de comprensión de las profundidades de la psique de una persona. Sin embargo, en la práctica, la interpretación en el Sistema integral de Exner se basa en un método muy empírico con referencia a una norma que sigue un procedimiento de codificación detallada. Las respuestas que presentamos y las de las otras manchas se codifican de manera cuidadosa; entonces se determinan ciertas sumas, porcentajes y razones de respuestas codificadas. Por último, estas sumas y razones se comparan con normas basadas en las respuestas de muestras de pacientes y no pacientes. En este sentido, interpretar las respuestas a las manchas de tinta del Rorschach se parece al procedimiento de criterio meta que revisamos al describir el MMPI. Por ejemplo, puede ser que el "atropellamiento" en la parte central de esta mancha sea bastante común en grupos de no pacientes. Por otro lado, puede ser que ver agua fluyendo en el borde de la mancha no sea común entre grupos de no pacientes, pero sí en personas que sufren de un estrés familiar inusual. ¿Cómo lo podríamos saber? Éste es el tipo de resultado que se obtiene al llevar a cabo numerosos estudios sobre la manera en que grupos bien definidos responden en realidad a las manchas de tinta. De la simple idea de "atropellamiento" no podemos concluir nada. Todas las respuestas deben codificarse. Si el examinado ve animales muertos, aplastados en la mayoría de las láminas, eso puede significar algo.

Sistema de codificación

Aquí ofrecemos un esbozo amplio de detalles selectos del proceso de codificación del Sistema integral. Al describir el MMPI-2 en el capítulo anterior, señalamos que esta prueba tiene un lenguaje propio; esto es aún más cierto en el caso del Rorschach, como será evidente cuando introduzcamos el sistema de codificación. En primer lugar, podemos observar que, aunque no se trata como categoría de codificación, el número total de respuestas (*R*) es el hecho inicial que se determina en el protocolo del Rorschach.

Exner (2003) hace hincapié en que al menos 14 respuestas deben estar presentes en el protocolo para que éste se considere interpretable. En el cuadro 14-2 aparecen ejemplos de las principales categorías de codificación del sistema de Exner. No se trata de listas completas, pues éstas han cambiado un poco con el paso del tiempo. En la parte superior del cuadro, enumeramos las categorías primarias de codificación.[2] En su mayor parte, estas puntuaciones son simples conteos de respuestas dentro de los códigos (más adelante describimos ejemplos de los códigos). Después, enumeramos razones, porcentajes y derivaciones, que son transformaciones de los conteos de las categorías primarias. Al final se encuentran las constelaciones. En general, una constelación es un indicador sí/no (p. ej., positivo de depresión) si se cumple cierto número de condiciones; éstas surgen de los reactivos que se incluyen en las categorías primarias de codificación o en las razones, porcentajes o derivaciones. Es evidente que el sistema de codificación es complejo y da origen a muchas puntuaciones. Ilustraremos algunos códigos y derivaciones para dar idea de la esencia del sistema.

Cuadro 14-2. Ejemplos de puntuaciones en el Sistema integral de Exner

Categorías primarias	
Ubicación	Contenidos
Calidad del desarrollo	Elementos populares
Determinantes	Puntuaciones especiales
Calidad de la forma	
Razones, porcentajes, derivaciones	
Ideación	Procesamiento
Afecto	Interpersonal
Mediación	Autopercepción
Constelaciones	
Índice de depresión	
Índice de pensamiento perceptual	
Índice de déficit de afrontamiento	
Índice de hipervigilancia	

El código más sencillo es el **código de ubicación**, que indica la zona de la lámina a la que el examinado hace referencia. El cuadro 14-3 muestra los códigos de ubicación.

Cuadro 14-3. Códigos de ubicación del Sistema integral del Rorschach

Símbolo	**Definición**	**Criterio**
W	Respuesta completa	La mancha entera se usa en la respuesta Todas las partes pueden usarse
D	Detalle común	Área de la mancha que se identifica con frecuencia

Dd	Detalle inusual	Área de la mancha que se identifica con poca frecuencia
S	Espacio	Área en blanco que se usa en la respuesta (se califica sólo con el símbolo de otra ubicación como en WS, DS o DdS)

Fuente: J. E. Exner. The Rorschach: A comprehensive system. Volume 1: Basic foundations (4th ed.), p. 77. Copyright © 2003. Reproducido con autorización de John Wiley & Sons, Inc.

Podemos notar dos características en los códigos de ubicación. Primero, podemos suponer que los códigos se referirán a ubicaciones específicas en la lámina (p. ej., parte izquierda superior, central, etc.), pero no es así. La principal diferencia en estos códigos es si la respuesta tiene que ver con la mancha completa o con una parte de ella. Segundo, la distinción entre D y Dd es exclusivamente empírica. Por ejemplo, en el caso de la mancha de la figura 14-1, si muchas personas se refieren a las extensiones superiores, esa respuesta se codifica como D; si sólo algunas personas hacen referencia a estas extensiones, la respuesta se codifica como Dd sin que esté implicada teoría alguna. ¿Cómo se define "muchas personas" (D) o "algunas personas" (Dd)? En el sistema de Exner, un detalle mencionado por *menos de 5%* de los examinados se considera poco común. La frecuencia de varias respuestas se determina en relación con los cuadros de normas del primer volumen de Exner.

Los códigos de los determinantes son los más complejos y extensos. Existen nueve categorías principales y 24 subcategorías de los determinantes. En el cuadro 14-4 aparece una lista de los códigos de una categoría principal y sus subcategorías para ilustrar estos tipos de códigos. Podemos notar qué tan específicas son las instrucciones para usar los códigos. En general, los **determinantes** indican qué características de la mancha influyeron o determinaron las respuestas del examinado. Por ejemplo, ¿en qué medida fueron determinantes el color percibido, el movimiento, las figuras humanas o animales, etc. en las otras subcategorías?

Cuadro 14-4. Ejemplos de categorías para codificar los determinantes en el sistema integral del Rorschach

Categoría	**Símbolo**	**Criterios**
Movimiento	M	*Respuesta de movimientos humanos*. Se usa en respuestas que implican actividad cinestésica de un humano, o de un personaje animal o ficticio en una actividad de tipo humana.
	FM	*Respuesta de movimientos animales*. Se usa en respuestas que implican una actividad cinestésica de un animal. El movimiento percibido debe ser congruente con la especie que se identifica en el contenido. Cuando el movimiento del animal no es común para la especie, se debe codificar como M.
	m	*Respuesta de movimientos inanimados*. Se usa en respuestas que implican movimiento de objetos inanimados, inorgánicos o absurdos.

Fuente: J. E. Exner. The Rorschach: A comprehensive system. Volume 1: Basic foundations (4th ed.), p. 87. Copyright © 2003. Reproducido con autorización de John Wiley & Sons, Inc.

Las respuestas **populares** son las que ocurren en al menos una tercera parte de las respuestas a una lámina en el grupo de estandarización. Se trata de una codificación estrictamente con referencia a una norma. Las 15 categorías de **Puntuaciones especiales** incluyen áreas como respuestas Mórbidas (MOR; hace referencia a la muerte, asesinato, etc.), Agresivas (Ag; pelear, atacar, etc.) y Movimiento cooperativo (COP). Estos códigos identifican características inusuales en las respuestas y se aplican sólo cuando éstas se encuentran presentes además de los códigos estándar de aplicación universal como la ubicación y los determinantes.

Como señalamos antes, las puntuaciones en las categorías de razones, porcentajes y derivaciones surgen de combinaciones de puntuaciones de los códigos primarios. Algunas de éstas son muy sencillas, por ejemplo, el porcentaje de respuestas que son "populares"; otras son más complejas, por ejemplo, agregar varios códigos con pesos especiales.

¡Inténtalo!

Regresa a las respuestas a la mancha de la figura 14-1 que aparecen en la página 366a». ¿Cómo aplicarías estos códigos de ubicación? Observa primero que en realidad hay dos respuestas (R = 2) que proporcionó el cliente.

Secuencia de puntuaciones y Resumen estructural

Los códigos ya descritos se resumen en la Secuencia de puntuaciones y el Resumen estructural. En la primera sólo se enumera el número de láminas, el número de respuestas y los códigos de cada respuesta, lo cual es importante para capturar los datos en la computadora y para fines de investigación. Tiene cierto potencial para hacer interpretaciones, pero no es un medio interpretativo importante.

El **Resumen estructural** es la fuente primaria para la interpretación, pues contiene resúmenes de todos los códigos además de una sección de razones, porcentajes y derivaciones que se obtienen a partir de los códigos. Por ejemplo, una entrada muestra la razón de W: M, es decir, la razón de todas las respuestas completas de los códigos de ubicación y las respuestas de Movimiento humano de los códigos de determinantes. El Resumen estructural, en esencia, no se puede interpretar si no se cuenta con una preparación avanzada en el Sistema integral de Exner. Los informes narrativos generados por computadora, semejantes a los que examinamos en el caso del MMPI-2 y otras pruebas, ahora están disponibles para el Rorschach.

Aunque el Sistema integral de Exner hace hincapié en establecer relaciones empíricas entre características personales y códigos específicos (y los derivados de ellos), existen algunos temas generales. Por ejemplo, las respuestas de "forma" son de especial relevancia para investigar la ideación; las respuestas de "color" se relacionan más con los estados emocionales.

Evaluación del Rorschach

La literatura de investigación sobre el Rorschach es tan vasta que casi desafía la posibilidad de resumirla. Leer incluso una muestra de esta literatura es casi como contestar un Rorschach: ¡lo que vemos puede estar determinado por predisposiciones personales más que por lo que realmente hay ahí! Por otro lado, podemos encontrar condenas absolutas del Rorschach; por ejemplo, Hunsley y Bailey (1999) concluyeron que "en la actualidad no existen bases científicas para justificar el uso de las escalas Rorschach en la evaluación psicológica" (p. 266). En Dawes (1994) y Lilienfeld *et al*. (2000) se puede encontrar un tratamiento muy poco amable sobre el Rorschach y, en general, de las técnicas proyectivas. Por otro lado, Weiner (2001) replicó que "la excesiva acusación de Hunsley y Bailey en contra del Rorschach como instrumento de evaluación por no cumplir con los estándares profesionales de la práctica ignora la abundante evidencia de lo contrario y carece de justificación" (p. 428). Viglione (1999) concluyó: "La evidencia revela que muchas variables del Rorschach son herramientas eficientes en aplicaciones clínicas, forenses y educativas" (p. 251). Viglione y Hilsenroth (2001) afirmaron que "una gran cantidad de evidencias empíricas apoya la confiabilidad, validez y utilidad del Rorschach y revela que las recientes críticas contra esta prueba, en gran medida, carecen de justificación" (p. 452). Con base en metaanálisis que comparan los coeficientes de validez de 31 estudios sobre el MMPI y 34 sobre el Rorschach, Hiller *et al*. (1999) concluyeron que la validez fue aproximadamente igual en el caso de estos dos instrumentos bien conocidos, aunque en ambos casos los coeficientes de validez, en promedio, fueron muy modestos. Atkinson (1986) y Parker, Hanson y Hunsley (1988) también concluyeron que el Rorschach era casi equivalente al MMPI en términos de validez. No hace falta decir que el debate sobre el Rorschach es intenso. Los interesados en saber más del Rorschach pueden consultar a Groth-Marnat (2009), donde encontrarán un capítulo dedicado a la revisión minuciosa del Rorschach, que incluye una excelente cronología de las varias etapas en la elaboración de la prueba. Weiner y Greene (2008) también resumieron la investigación actual (y las controversias) alrededor del Rorschach. El metaanálisis de Mihura *et al*. (2013) también es útil.

¿Qué sentido podemos darle a estas conclusiones contradictorias? Para ofrecer cierta perspectiva, señalamos que, igual que en el caso de cualquier otra prueba, debemos concentrarnos en puntuaciones específicas y no en la prueba como técnica general. Con esto en mente, podemos aventurar las siguientes generalizaciones. Primero, la aplicación del sistema de Exner da por resultado la generación de puntuaciones confiables de muchas variables, pero no todas. La confiabilidad interjueces, basada en usuarios con una cuidadosa preparación, es bastante buena en el caso de muchas puntuaciones. La confiabilidad de test-retest varía ampliamente en distintas puntuaciones, pero parte de esta variabilidad se debe, quizá, a cambios reales a corto plazo en el rasgo evaluado por una puntuación particular. La evidencia de la validez muestra diferencias notables de una fuente a otra. Numerosos estudios han obtenido una validez significativa en varias

puntuaciones del Rorschach, pero otros, también numerosos, han salido con las manos vacías al intentar demostrar la validez de esta prueba. En general, concluimos que, cuando la prueba se aplica y califica de manera estandarizada, sin duda puede producir evidencia de una validez respetable.

¿Y ahora qué sigue?: llega el R-PAS

La influencia de John Exner en el Rorschach es impresionante, pero él murió en 2006. ¿Y ahora qué sigue? Al parecer, la respuesta vendrá de un equipo integrado, en parte, por personas asociadas estrechamente con Exner: Meyer, Viglione, Mihura, Erard y Erdberg (2011), con su recién lanzado *Rorschach Performance Assessment System* [Sistema de evaluación del desempeño en el Rorschach], R-PAS. Este sistema busca promover el desarrollo psicométrico de la interpretación del Rorschach, extender los esquemas actuales de codificación, mejorar los sistemas de calificación y darle un toque internacional a todo esto. ¿El R-PAS representará el nuevo sistema integral? Estaremos al pendiente.

Test de Apercepción Temática (TAT)

El Test de Apercepción Temática (**TAT**; Murray, 1943) es la segunda técnica proyectiva más usada de acuerdo con distintas encuestas (Camara *et al*., 2000; Piotrowski & Keller, 1984, 1989; Wade & Baker, 1977). Por muchos años se ha mantenido entre las 10 pruebas más usadas (Lubin, Larsen, & Matarazzo, 1984). En términos de la importancia para la formación de psicólogos clínicos, el TAT superó al Rorschach en la categoría de pruebas proyectivas; sin embargo, existe evidencia de que su uso ha disminuido en años recientes.

El TAT consta de 30 láminas, 29 de las cuales contienen un dibujo y una está en blanco. Cada lámina mide 23 × 30 cm, y el dibujo cubre casi dos terceras partes de ella. La figura 14-2 muestra un dibujo similar a los que se usan en el TAT. En esta prueba se eligieron deliberadamente dibujos ambiguos, aunque no tanto como una mancha de tinta.

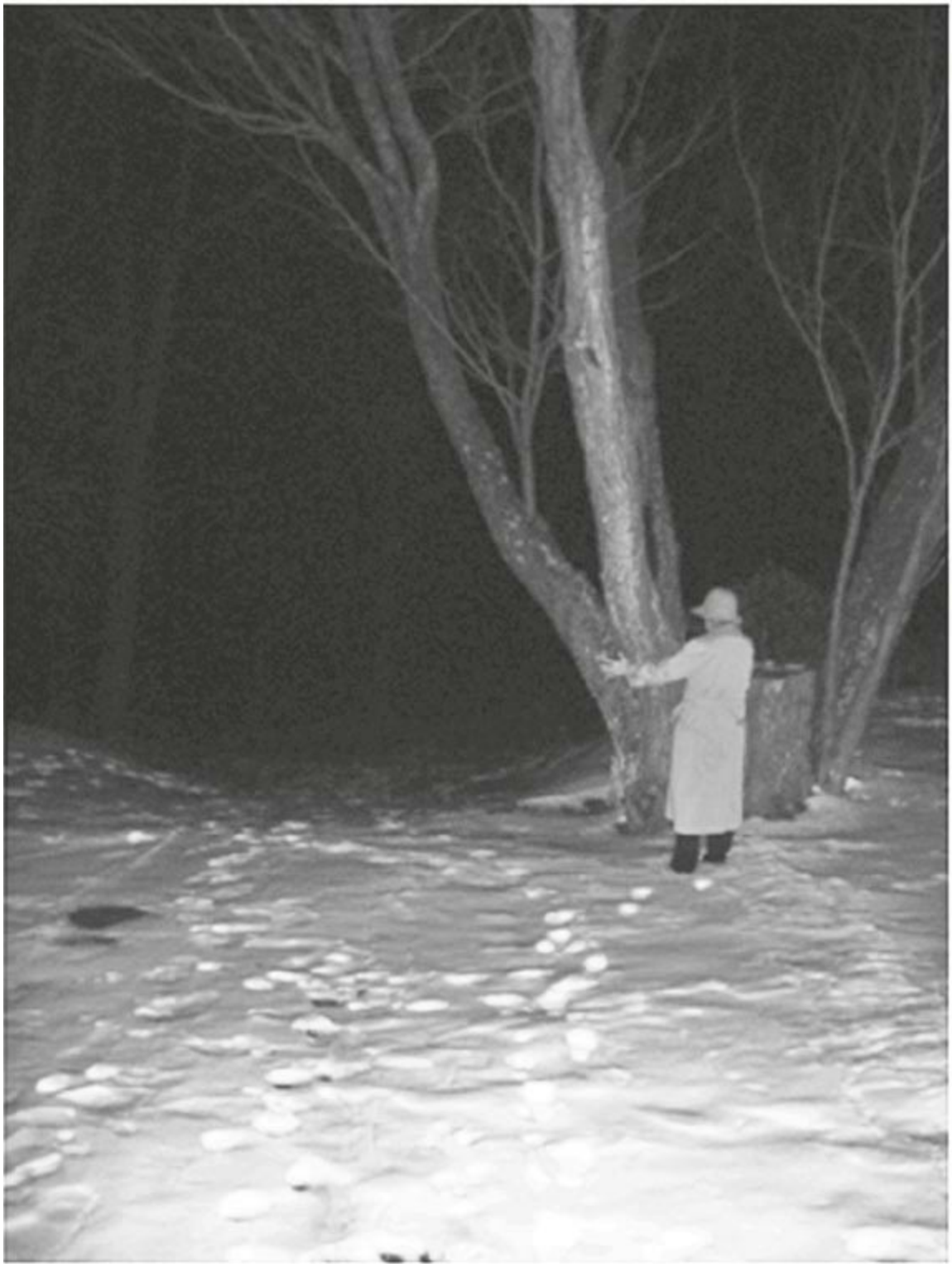


Figura 14-2. Imagen similar a las que se usan en el TAT.

No todas las láminas se usan con todos los examinados; la selección depende de si se trata de adolescentes o adultos, o de mujeres u hombres. Once láminas, incluida la que está en blanco, se pueden aplicar a cualquier examinado, siete son sólo para adolescentes

y adultos varones, siete para mujeres adolescentes o adultas y una para cada uno de los subconjuntos, por ejemplo, mujeres adultas. Un código en la parte posterior de la lámina indica a qué grupo está dirigida, por ejemplo, la lámina 12VN está dirigida a adolescentes varones y mujeres, mientras que la lámina 12H está dirigida sólo a hombres adultos.

Doce láminas del TAT presentan a una persona, 11 tienen dos personas, cinco tienen más de dos personas y en dos no aparece ninguna persona. La mayoría de los dibujos tiene una apariencia misteriosa, algunos dirían lúgubre. (¡Se debe tener cuidado al caracterizar los dibujos para que la descripción no se tome como una respuesta proyectiva!)

El origen del TAT se encuentra en el trabajo de Henry Murray (Murray *et al*., 1938), quien postuló un conjunto de necesidades psicológicas (afiliación, autonomía, agresión, etc.) y presiones (fuerzas ambientales). Murray pensaba que las respuestas a los dibujos ambiguos del TAT ayudarían a revelar las necesidades y presiones dominantes de una persona.

Las instrucciones originales para aplicar el TAT fueron las siguientes (Murray *et al*., 1938, p. 532):

> El sujeto se sienta en una silla cómoda de espaldas al experimentador y se le leen las siguientes instrucciones: Ésta es una prueba de su imaginación creativa. Le mostraré un dibujo y quiero que invente una trama o historia en la que éste pueda usarse como ilustración. ¿Qué relación hay entre los individuos del dibujo? ¿Qué les ha ocurrido? ¿Cuáles son sus pensamientos y sentimientos actuales? ¿Cuál será el resultado? Intente hacerlo lo mejor que pueda. En vista de que le pido dar rienda suelta a su imaginación literaria, su historia puede ser tan larga y detallada como usted lo desee.

En el esquema original de Murray, cada examinado respondía a 20 láminas. El tiempo promedio por lámina era de 5 min, por lo que se requerían casi 2 hrs que debían dividirse en dos sesiones de evaluación.

Desafortunadamente, las disposiciones de aplicación del TAT son muy variadas en la práctica. Primero, mientras que la aplicación completa exige 20 láminas, lo cual requiere dos sesiones de una hora, casi nadie usa en la actualidad las 20 láminas con un examinado, por lo que es difícil comparar los resultados de distintos examinados. Segundo, las instrucciones exactas pueden variar de un examinador a otro; al parecer, la mayoría emplea instrucciones parecidas a las de Murray que presentamos antes, pero aún existen notables variaciones en la práctica. Muchos usuarios encuentran engañosa la frase de que es una prueba de imaginación creativa y, por lo tanto, no la usan. En casi todas las situaciones, el examinador pide una historia que cuente lo que está sucediendo, qué llevo a esto y qué podría suceder después. Cuando es necesario, se alienta al examinado a identificar los pensamientos y sentimientos implicados en el dibujo. Al igual que en el Rorschach, es preferible más que menos en las respuestas. Por último, mientras que el Sistema integral de Exner se convirtió en la aplicación y calificación estándar del Rorschach, en el caso del TAT no existe un sistema comparable.

La obra contemporánea sobre el TAT más influyente es la de Leopold Bellak. Su libro, ahora en la sexta edición (Bellak & Abrams, 1997), constituye el "verdadero" manual del

TAT para muchos usuarios. Bellak ha intentado hacer para el TAT lo que Exner hizo para el Rorschach: sistematizar la aplicación y calificación, y llevar a cabo la investigación psicométrica necesaria. Bellak recomienda el uso exacto de 10 láminas (1, 2, 3VH, 4, 6VH, 7NM, 8VH, 9NM, 10 y 13HM). El sistema de calificación de Bellak tiene categorías específicas, que incluyen variables como tema principal, héroe principal y necesidades e impulsos básicos del héroe. Estas categorías contrastan con un uso basado en impresiones de las respuestas. El método de Bellak tiene un toque distintivo psicoanalítico; al mismo tiempo, adopta descaradamente un duro ataque psicométrico en cuestiones de confiabilidad, validez y normas. El libro de referencia de Bellak también trata dos derivados del TAT: el Test de Apercepción Infantil (CAT) y el Test de Apercepción Temática para Edades Avanzadas (SAT). Quienes se interesen con seriedad en el TAT deben consultar la obra de Bellak. Jenkins (2008) compiló numerosos sistemas para calificar el TAT.

Las respuestas del TAT pueden ser escritas en vez de orales. Las respuestas orales son la norma en el uso clínico. Las respuestas escritas a veces se utilizan para fines de investigación; tienen la obvia ventaja de permitir la aplicación grupal y el uso de muestras de investigación más grandes. Sin embargo, existen diferencias sistemáticas entre las respuestas orales y escritas (véase, p. ej., Dana, 1996).

¿Qué dice la investigación acerca de las características psicométricas del TAT? Debido a la diversidad de procedimientos de aplicación y calificación, es mucho más difícil hacer generalizaciones sobre el TAT que sobre el Rorschach o el RISB, que veremos en la siguiente sección. Los investigadores han establecido una confiabilidad y validez respetables cuando se usa el TAT con constructos bien definidos. El trabajo de McClelland y Atkinson con los constructos de motivación de logro y afiliación ilustra esta generalización (véase Atkinson, 1958; McClelland, 1985; McClelland, Atkinson, Clark, & Lowell, 1953; Spangler, 1992). Sin embargo, estos resultados se han obtenido, en su mayoría, en escenarios atípicos de práctica clínica, que constituye el sitio primario de aplicación del TAT. En general, los autores de reseñas no han sido amables con el TAT; aunque señalan que el TAT es una fuente potencialmente rica de información, casi todos censuran el hecho de que su aplicación y calificación no esté estandarizada.

¿Qué podemos aprender de nuestro examen del TAT? Primero, obtener respuestas a dibujos ambiguos ha demostrado ser un dispositivo muy popular entre los psicólogos. Al parecer, existe un sentimiento muy arraigado de que esta técnica permite que la hipótesis proyectiva funcione. La popularidad del TAT por tantos años es sobresaliente. Segundo, observamos que el uso del TAT parece disminuir, ¿por qué? Es casi seguro que esto se pueda atribuir a la falta de un sistema sistemático de calificación (Dana, 1996). Los psicólogos han prestado atención cada vez más a la necesidad de confiabilidad, validez y normas en las pruebas. La ausencia de un sistema de calificación dominante y definido con claridad hace difícil reunir datos para obtener la confiabilidad, validez y normas necesarias en el uso contemporáneo de las pruebas. La falta de estandarización probablemente continuará desplazando a esta prueba, lenta pero inexorablemente, hacia la periferia del campo de las pruebas. En Groth-Marnat (2009) y Weiner y Greene (2008)

se puede encontrar discusiones del uso clínico del TAT, reseñas de la investigación y lamentaciones continuas por su falta de estandarización.

Frases Incompletas de Rotter (RISB)

Como señalamos antes en este capítulo, hay muchos ejemplos de las pruebas de frases incompletas: una revisión reciente de la base de datos de ETS Test Collection reveló más de 30. Una de ellas destaca en términos de la frecuencia con que se usa, la extensión de su base de investigación y su reputación general; nos referimos al *Test de Frases Incompletas de Rotter* [Rotter Incomplete Sentences Blank], Segunda Edición (**RISB**; Rotter, Lah, & Rafferty, 1992). Holaday *et al*. (2000) ofrecen una breve descripción de 15 distintas pruebas de frases incompletas (PFI) y encuestaron a miembros de la Society for Personality Assessment [Sociedad para la Evaluación de la Personalidad] respecto del uso de dichas pruebas. El RISB fue con claridad la opción más popular, mientras que el resto de las pruebas tuvo muy pocos usuarios. La edición actual del RISB difiere en grado mínimo de la primera (Rotter & Rafferty, 1950) en términos de propósito, estructura y troncos de oraciones; sin embargo, el manual de la prueba se ha actualizado de manera minuciosa, en especial resumiendo la vasta investigación que se ha llevado a cabo con el RISB en los 40 años que transcurrieron entre la primera y segunda edición. El manual también incluye guías mejoradas de calificación y nuevas normas.

Existen tres formas del RISB: bachillerato, universidad y adultos. (El RISB usa el término *forma* para referirse a diferentes niveles de la prueba más bien que en el sentido de formas alternas más o menos equivalentes.) La forma universitaria fue la original y es la que más se ha empleado para hacer investigaciones. Las otras dos formas difieren de la universitaria sólo en algunos troncos de reactivos. Regresaremos a la cuestión de las formas más adelante; mientras tanto, a menos que digamos lo contrario, discutiremos la forma universitaria.

El RISB consta de 40 frases incompletas o *troncos*. Los troncos, por lo común, constan de sólo dos palabras, a veces sólo una, aunque hay troncos de cuatro o cinco palabras. El cuadro 14-5 muestra troncos como los que aparecen en el RISB y otras PFI; como se muestra aquí, las hojas de la prueba proporcionan sólo una línea por cada reactivo para registrar las respuestas del examinado. El total de reactivos y los espacios para las respuestas caben en las dos caras de una hoja de 21.5 × 28 cm.

Cuadro 14-5. Troncos de reactivos típicos de las pruebas de frases incompletas

Espero _____
La gente suele _____

Las instrucciones son muy sencillas: se pide al examinado que exprese sus "verdaderos sentimientos" completando la oración en cada reactivo. El examinado, por lo común, termina la prueba en 20 o 25 min.

El RISB difiere de otras técnicas proyectivas usadas con frecuencia, por ejemplo, el Rorschach y el TAT, de dos maneras importantes. Primero, la disposición física de la prueba promueve una respuesta concisa a cada reactivo. Es posible que un examinado

divague al completar una oración, pero es poco probable que suceda; en contraste, en el Rorschach y el TAT se anima al examinado a extenderse en sus respuestas.

Segundo, el RISB pretende medir sólo un constructo: *adaptación* (o su contrario, desadaptación). De manera congruente con este propósito único, el RISB produce sólo una puntuación: Adaptación general. El manual de esta prueba define **adaptación** como:

> la relativa libertad respecto de estados (emociones) prolongados de infelicidad/disforia, la capacidad para afrontar la frustración, la capacidad para iniciar y mantener actividades constructivas y la capacidad para establecer y mantener relaciones interpersonales satisfactorias. (Rotter *et al*., 1992, p. 4)

Desde luego, la desadaptación es lo contrario de estas características.

Cada reactivo del RISB se califica de acuerdo con los signos de adaptación/desadaptación en una escala de 7 puntos (0-6), en la cual 6 indica la desadaptación más grave, 0 indica una respuesta saludable muy positiva y 3 representa una respuesta "neutral" que no significa ni adaptación buena ni mala. La suma de la puntuación de cada reactivo se prorratea si hay respuestas omitidas. La suma de todos los reactivos constituye la puntuación de Adaptación general, que puede variar de 0 (muy bien adaptado) a 240 (sumamente desadaptado, es decir, cada reactivo tuvo una puntuación de 6). En los grupos de estandarización, la mayoría de las puntuaciones de Adaptación general se ubica en el rango de 100 a 170. Podemos observar que las respuestas "neutrales" a todos los reactivos producirían una puntuación total de $3 \times 40 = 120$. La media de los grupos de estandarización (véase más adelante) es casi de 130.

¡Inténtalo!

Completa los dos troncos de reactivos que aparecen en el cuadro 14-5, de modo que, en tu consideración, las respuestas indiquen un grado moderado de adaptación, es decir, que se calificarían con 2 en la escala de 7 puntos. Después, escribe respuestas que indiquen un grado moderado de desadaptación, es decir, que se calificarían con 5.

El manual del RISB ofrece indicaciones detalladas para calificar los reactivos. Presenta la lógica general para la calificación, respuestas muestras para cada puntuación de cada reactivo (de manera separada para hombres y mujeres) y un apéndice de seis casos para practicar la calificación. En este aspecto, el manual del RISB recuerda los manuales de pruebas individuales de inteligencia como las escalas Wechsler.

El manual del RISB resume numerosos estudios de la confiabilidad y validez de la puntuación de Adaptación general. La confiabilidad interjueces, desde luego, como sucede con otras medidas proyectivas, es tema de especial preocupación. Varios estudios que se resumen en el manual documentan que, de hecho, la prueba tiene una buena confiabilidad de interjueces, en promedio, alrededor de .90. La confiabilidad de consistencia interna (división por mitades y alpha de Cronbach) tiene un promedio cercano a .80. Los coeficientes de confiabilidad de test-retest varían mucho dependiendo

del intervalo entre la primera y la segunda aplicación. En el caso de intervalos de 1 o 2 semanas, estos coeficientes son cercanos a los de consistencia interna, mientras que en el de intervalos de varios meses o años, los coeficientes de desploman hasta llegar a .50 en promedio. Esta situación nos recuerda la confiabilidad del *Inventario de Ansiedad Rasgo-Estado* (véase capítulo 13). Parece que la adaptación, como la mide el RISB, es un *estado* razonablemente estable, pero no un *rasgo* estable. El manual del RISB contiene una discusión franca sobre esta cuestión.

Se ha llevado a cabo una gran cantidad de estudios de validez del RISB, de los cuales el manual presenta una revisión completa. Se incluyen investigaciones en las que se hacen contrastes entre grupos conocidos, correlaciones con otras medidas de adaptación, ansiedad y otros constructos de la personalidad, análisis factoriales y relaciones con la inteligencia y el aprovechamiento. En general, los estudios de validez apoyan la idea de que el RISB mide el constructo de adaptación, aunque, como suele suceder con las pruebas psicológicas, la evidencia de estos estudios no es definitiva. La fortaleza del manual del RISB es que presenta muchos estudios, con lo cual brinda al usuario un contexto adecuado para llegar a sus propias conclusiones.

Derivado de un estudio de validez con grupos contrastados que se informa en el manual, se sugiere 145 como puntuación de corte para identificar casos de desadaptación. Sin embargo, la base de este contraste –autorreferencias a un centro de orientación y una muestra general de universitarios– no es sólida. Esta puntuación de corte equivale aproximadamente a una desviación estándar arriba de la media de los grupos de estandarización.

El manual del RISB presenta normas por género basadas en 110 mujeres y 186 hombres de tres estudios que se llevaron a cabo con distintas muestras de 1977 a 1989. El manual ofrece poca información acerca de la naturaleza de estas muestras; además, no se presentan normas para las formas de bachillerato y adultos de la prueba. Por tratarse de una prueba muy usada, esperaríamos más: el número de casos e información sobre los casos de los grupos de estandarización. El manual muestra las medias y desviaciones estándar de una gran cantidad de estudios publicados, pero esto no es un sustituto adecuado de un conjunto decoroso de normas. Incluso las que aparecen son simples porcentajes acumulados de intervalos de 5 puntos de puntuaciones de Adaptación general. El manual recomienda desarrollar normas locales; se trata de una práctica útil sólo si la definición de adaptación/desadaptación difiere de manera considerable de una población local a otra, lo cual es dudoso.

Una de las características peculiares del RISB como técnica proyectiva es que se propone con firmeza medir sólo una variable, adaptación, mientras que la mayoría de estas técnicas pretende medir muchas. ¿El RISB se puede usar para medir algo diferente de la adaptación? Sin duda, varias personas han tratado de hacerlo; el manual del RISB identifica 15 estudios que intentaron usar la prueba para medir docenas de otras variables (p. ej., ansiedad ante la muerte, dependencia, hostilidad, etc.). El manual toma distancia de estos estudios; no censura ni respalda los esfuerzos por medir otras variables, sino sólo señala que se han hecho.

¿Qué podemos aprender al examinar el RISB? Al menos, podemos citar las siguientes lecciones. Primero, es claro que podemos desarrollar criterios muy específicos de calificación para un estímulo proyectivo.

Segundo, se puede lograr una buena confiabilidad interjueces con estos criterios. Las primeras dos generalizaciones son, sin duda, similares a las lecciones que aprendimos al examinar el trabajo de Exner con el Rorschach. Tercero, para cumplir los dos últimos objetivos, es importante tener un constructo claro en mente; en el caso del RISB, el constructo es adaptación/desadaptación. Cuarto, este rasgo (adaptación), tal como lo mide el RISB, parece ser estable en períodos cortos, pero no en períodos mayores de algunos meses. Aunque el rasgo no es por completo inestable, como lo sería un castillo de arena frente a la marea alta, no evidencia una estabilidad a largo plazo como la inteligencia verbal. Por último, observamos la necesidad de normas bien establecidas y puntuaciones de corte para pruebas como ésta.

Se puede obtener más información sobre el RISB, Segunda Edición, en Boyle (1995) y McLellan (1995). Weiner y Greene (2008) resumieron el uso y la investigación sobre el RISB y el *Washington University Sentence Completion Test* [Prueba de Frases Incompletas de la Universidad de Washington]. Holaday *et al*. (2000) proporcionaron información útil sobre el uso normal del RISB por parte de los clínicos.

Dibujos de la figura humana

Quizá más que cualquier otra área de las pruebas, la de los dibujos de figuras humanas está poblada de iniciales, a veces de una manera muy confusa (cuadro 14-6). En el nivel más específico, tenemos el **HTP**, Prueba Casa-Árbol-Persona [*House-Tree-Person*], y el **KFD**, Prueba de Dibujo Cinético de la Familia [*Kinetic Family Drawing Test*], que describiremos más adelante. Después se encuentra el **DAP**, Draw-A-Person [Dibuja a una Persona], que a menudo es ambiguo. Las iniciales pueden representar una prueba específica o pueden abarcar todas las pruebas en que se pide dibujar personas. En este sentido, el DAP incluye el *Draw-A-Man Test* [Prueba Dibuja a un Hombre], precursor de todas estas pruebas, que por cierto ¡nunca se abrevia como DAM! Por último, encontramos el **DFH**, Dibujo de la Figura Humana. A veces se designa de esta manera a todos los dibujos proyectivos que incluyen humanos, es decir, DAP, HTP, KFD y otras variantes. En otras ocasiones, DFH es equivalente al DAP, es decir, incluye todos los dibujos de personas; y en otras más, DFH hace referencia a una prueba específica de Draw-A-Person.

Cuadro 14-6. Abreviaturas comunes que representan dibujos proyectivos

DFH	Dibujo de la Figura Humana
DAP o D-A-P	Draw-A-Person
HTP o H-T-P	Prueba Casa-Árbol-Persona
KFD	Prueba de Dibujo Cinético de la Familia

Los dibujos de la figura humana, por lo regular, se encuentran entre las pruebas más usadas; pero antes de comentar sobre la frecuencia de su uso, debemos reiterar un punto que ya mencionamos antes. Existen distintas pruebas que implican dibujar figuras humanas, las cuales constituyen una técnica general, no una prueba específica. Sin embargo, muchas encuestas sobre uso de pruebas agrupan todos los métodos en la categoría de "dibujos de la figura humana" o "dibujo de personas", de modo que no se puede saber con exactitud qué pruebas se usan.

Además, los psicólogos emplean los dibujos de la figura humana para un rango sumamente amplio de propósitos, que incluyen la evaluación de la personalidad, la inteligencia y la disfunción neuropsicológica. Estos factores complican nuestra comprensión de uso habitual de estos dibujos.

Con estas precauciones en mente, señalamos los siguientes resultados de los dibujos de la figura humana en las encuestas sobre el uso de pruebas. Entre psicólogos clínicos, orientadores y escolares, los dibujos de la figura humana se encuentran entre las pruebas usadas con mayor frecuencia (véase Archer *et al*., 1991; Camara *et al*., 2000; Hutton *et al*., 1992; Kennedy *et al*., 1994; Piotrowski & Keller, 1984). Cuando las pruebas se identifican por separado, el DAP suele ser el más usado, seguido de cerca por el HTP. El DAP suele encontrarse después del Rorschach y el TAT (y, a veces, frases incompletas)

en la frecuencia de su uso entre las pruebas proyectivas, aunque en algunos casos el DAP supera a estas dos pruebas.

¡Inténtalo!

Para poder apreciar la variedad de pruebas que implican dibujos de la figura humana, escribe las palabras clave "draw a person" en el sitio de internet de ETS Test Collection (http://www.ets.org/test_link/find_tests/) o en una base de datos como PsychINFO. Observa el número de distintas pruebas que resultan de la búsqueda con estas palabras clave.

La primera prueba de este tipo que se utilizó fue la de Florence Goodenough (1926), Draw-A-Man Test, que pretendía ser una medida de inteligencia no verbal. A través de sus sucesores (Harris, 1963; Naglieri, 1988), este método aún se usa como medida de inteligencia. Sin embargo, no mucho después de que se introdujo esta prueba, varios psicólogos empezaron a usarla como medida proyectiva de la personalidad. El esfuerzo más famoso y sistemático para usar con dicho propósito estos dibujos fue el de Machover (1949). En Naglieri, McNeish y Bardos (1991) se puede encontrar otro ejemplo de un intento de usar los dibujos de la figura humana como medida de la personalidad.

Las instrucciones para aplicar el DAP varían mucho, pero suelen contener los siguientes elementos clave. Se da a la persona una hoja blanca de 21.5 × 28 cm y un lápiz con goma para borrar, y se le pide dibujar a una persona. A veces, las instrucciones son "haga un dibujo de usted mismo", a veces sólo se dice "haga un dibujo de una persona". Cuando el examinado termina el dibujo, se le dice "dibuje a una persona del sexo opuesto". Los dibujos suelen hacerse en 5 o 10 min. El examinador puede hacer preguntas acerca de las características del dibujo cuando ya está terminado.

Existen varios sistemas para calificar los dibujos de la figura humana, pero ninguno detenta una posición dominante. Se podría decir que el trabajo histórico más influyente en estos dibujos es el de Machover (1949). Las afirmaciones de la obra de Machover se han vuelto parte del saber popular que rodea a estos dibujos. El cuadro 14-7 presenta una muestra de afirmaciones típicas de Machover, muchas de las cuales han permeado en otros trabajos. Las caracterizaciones como las que aparecen en el cuadro 14-7 se dan para prácticamente cada detalle del cuerpo. Las generalizaciones fáciles, que no tienen apoyo de ninguna investigación citada, abundan en el trabajo. Adoptadas por personas sin la suficiente sensibilidad a la necesidad de verificación empírica, las afirmaciones se han tomado como hechos más que como hipótesis que se deben explorar.

Cuadro 14-7. Afirmaciones de Machover (1949) sobre los dibujos de la figura humana
"Cabezas desproporcionadamente grandes aparecen a menudo en los dibujos de individuos que sufren una enfermedad orgánica cerebral, que se han sometido a cirugía cerebral o que han estado aquejados de dolores de cabeza o alguna sensibilidad cerebral especial." (p. 37)
"Los sujetos que omiten de manera deliberada características faciales en un dibujo, muestran una

delineación cuidadosa y a menudo agresiva del contorno y detallan otras partes de la figura, se muestran evasivos acerca de las fricciones en sus relaciones interpersonales." (p. 40)
"Ya que la boca es a menudo fuente de satisfacción sensual y erótica, destaca a primera vista en los dibujos de individuos con dificultades sexuales." (p. 43)
"Un cuello largo y a menudo delgado, que resulta en una separación llamativa del cuerpo respecto de la cabeza, casi siempre se observa en individuos esquizoides o, incluso, esquizofrénicos." (pp. 57-58)

Cuando se tratan como hipótesis y se ponen a prueba, los resultados han sido decepcionantes, por decir lo menos.

Varios autores han intentado establecer en términos psicométricos métodos sólidos para los procedimientos del DFH. Un buen ejemplo es *Draw-a-Person: Screening Procedure for Emotional Disturbance* [Dibuja-una-persona: Procedimiento exploratorio de perturbaciones emocionales] (DAP: SPED; Naglieri *et al.*, 1991). El resultado general parece ser un modesto éxito. En Cosden (1995) y Morrison (1995) se pueden encontrar revisiones del DAP: SPED. Se puede alcanzar cierto grado de confiabilidad, pero la validez parece estar limitada, en gran parte, a un efecto del tamaño moderado para identificar desadaptación general, mientras que distinciones más refinadas siguen siendo escurridizas.

Dos ramificaciones populares de la técnica dibuja-una-persona son *Casa-Árbol-Persona* (HTP; Buck, 1948, 1966) y el *Dibujo Cinético de la Familia* (KFD: Burns & Kaufman, 1970, 1972), las cuales están dirigidas principalmente a niños. En ambas pruebas, la teoría es que el niño revela elementos únicos, quizá inconscientes, de su personalidad por medio de estos dibujos con mayor probabilidad que sólo con dibujos de una persona. En el HTP, como lo sugiere su nombre, el niño dibuja un árbol, una casa y una persona, mientras que, en el KFD, hace el dibujo de una familia "haciendo algo". (*Cinético* deriva del griego *kinetikos* y significa relacionado con el movimiento o energía.) En una versión alterna se pide al niño que dibuje algo que esté ocurriendo en la escuela.

La esperanza es eterna respecto del potencial del DFH para revelar aspectos ocultos de la personalidad humana. En una revisión comprensiva, incluso proactiva, de varios procedimientos, Handler (1996) opina que, aunque la investigación hasta ahora ha sido decepcionante, quizá los investigadores no han realizado los estudios correctos. Groth-Marnat (1999) se refirió en un tono lacónico a la "investigación empírica, en gran medida, desalentadora" y a "que hacen falta muchas explicaciones de éstas tan inexplicablemente populares técnicas proyectivas" (p. 506); y en las subsiguientes ediciones de su tan citado libro (Groth-Marnat, 2003, 2009) dejó de incluir los dibujos de la figura humana. Weiner y Greene (2008) afirmaron que "las bases psicométricas de los métodos del dibujo de la figura humana son, en la actualidad, inciertas" (p. 510). Quizá esa evaluación es demasiado generosa; no es que la evidencia sea incierta, sino que es muy claramente desalentadora. Todos parecen estar de acuerdo en que la técnica puede ser útil para formular hipótesis y para romper el hielo.

Resumen de puntos clave 14-2

Factores que influyen en el uso futuro de las técnicas proyectivas

- Formación de psicólogos
- Atención administrada
- Demanda de puntuaciones objetivas, normas y calidad psicométrica

El futuro de las técnicas proyectivas

Como se evidencia en las distintas encuestas de uso de pruebas citadas a lo largo de este capítulo, las técnicas proyectivas se han afianzado con solidez en la práctica de la psicología. ¿Qué hay del futuro? ¿Se seguirán usando las pruebas proyectivas? ¿Habrá cambios en la manera en que se usan? Las respuestas a estas preguntas no están por completo claras, pero podemos, al menos, identificar *tres factores* pertinentes para responderlas (véase Resumen de puntos clave).

Formación de psicólogos

El primer factor se relaciona con la *formación de psicólogos*, en especial en áreas como la psicología clínica, de orientación y escolar. Es claro que los psicólogos se siguen formando en el uso de técnicas proyectivas. Al parecer, los responsables de su formación consideran importante ser eficiente en el uso de ellas. Belter y Piotrowski (2001) hicieron una encuesta sobre los programas doctorales aprobados por la APA en psicología clínica respecto de las pruebas que se estudiaban en los cursos. Encontraron que dos de las cinco pruebas presentes en más de la mitad de los programas eran técnicas proyectivas (el Rorschach y el TAT). Clemence y Handler (2001) encuestaron a directores de sitios de formación en la práctica para identificar qué técnicas de evaluación los directores consideraban importantes para que los internos las conocieran. Cinco de las 10 pruebas citadas eran técnicas proyectivas. (Las cinco se citan en el cuadro 14-1, al principio de este capítulo.) Además, los psicólogos que se están formando en la actualidad probablemente usen en la práctica las pruebas que emplean durante su formación. Sumando estos factores a la inercia propia de la práctica en cualquier campo, llegamos a la conclusión de que las técnicas proyectivas seguirán teniendo un papel destacado en la evaluación psicológica en el futuro inmediato. En Craig y Horowitz (1990), Culross y Nelson (1997), Hilsenroth y Handler (1995), Marlowe, Wetzler y Gibbings (1992), Piotrowski y Keller (1984) y Watkins, Campbell y McGregor (1988) se pueden encontrar prácticas y recomendaciones adicionales respecto de la formación de psicólogos en el uso de las técnicas proyectivas.

Surgimiento del manejo cuidadoso

Segundo, el *surgimiento del manejo cuidadoso* en los servicios de salud influye en el uso de pruebas (véase Acklin, 1996; Ben-Porath, 1997; Piotrowski, 1999). Esta tendencia puede tener implicaciones especiales para el uso de las técnicas proyectivas, pues el énfasis del manejo cuidadoso recae en el diagnóstico específico y el tratamiento inmediato más que en una valoración más global y resultados a largo plazo. Nietzel, Bernstein y Milich (1998) señalan sucintamente que “en la actualidad se pide a muchos

clínicos que limiten sus evaluaciones a la exploración inicial del paciente, la rápida medición de síntomas, el diagnóstico diferencial de los trastornos y las recomendaciones para el tratamiento" (p. 100). Las técnicas proyectivas parecen ser adecuadas en particular para análisis exploratorios en profundidad, ligados de manera natural con tratamientos más holísticos y sistémicos. Además, estas técnicas son intrínsecamente más caras que las pruebas objetivas de personalidad. Consideremos esta comparación. El cliente A llega a la clínica; un no profesional le entrega al cliente una prueba objetiva de personalidad y le explica de manera breve el propósito de la prueba y el procedimiento para responder las preguntas con un teclado. El cliente teclea sus respuestas. Tres minutos después, el cliente termina y el psicólogo clínico recibe el informe generado por computadora que contiene el perfil de puntuaciones y el comentario interpretativo. El psicólogo revisa el informe durante 10 min y, con una gran cantidad de información en la mano, está listo para entrevistar al cliente. El cliente B llega a la clínica. El psicólogo clínico aplica una prueba proyectiva, el Rorschach, durante una hora; después, dedica otra hora a calificar (codificar) las respuestas. Por último, se lleva otra media hora para interpretar las puntuaciones antes de entrevistar al cliente B. El tiempo profesional invertido en el cliente A fue de 10 min, mientras que en el Cliente B, 150 min. Tomando en cuenta esta diferencia, debe haber un fuerte argumento para justificar el uso de la técnica proyectiva. De hecho, cuando el costo es una preocupación, esta diferencia parece provocar que el uso de las técnicas proyectivas se reduzca.

Calificación objetiva, interpretación con referencia a una norma y calidad psicométrica

Tercero, cuando las técnicas proyectivas se utilizan de manera formal, parece claro que la tendencia es preferir la *calificación objetiva*, la *interpretación con referencia a una norma* y la *calidad psicométrica*. El ejemplo más claro de esta tendencia es la manera en que el Sistema integral de Exner ha influido en el uso del Rorschach. Como señalamos antes, este sistema es el estándar actual, aunque no sea el único que se usa (Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neulieb, 1999; Hilsenroth & Handler, 1995; Piotrowski, 1996). En general, el surgimiento del sistema de Exner ha sensibilizado a los usuarios de las técnicas proyectivas a la necesidad de instrucciones estandarizadas, codificación sistemática, confiabilidad adecuada y validez demostrada. El uso del TAT, cada vez más criticado debido a su falta de aplicación y calificación estandarizadas, parece ser una advertencia (Dana, 1996). Sin embargo, como señalamos antes, el trabajo de Bellak puede tener un efecto saludable en el uso del TAT. Las técnicas proyectivas elaboradas en años más recientes, así como los nuevos desarrollos de las técnicas antiguas, informan de manera cotidiana datos acerca de las normas, confiabilidad interjueces, validez concurrente y otros temas psicométricos. En el pasado, era común oír afirmaciones de que estos temas eran irrelevantes para las técnicas proyectivas, pero ahora prácticamente nadie hace este tipo de afirmaciones. Sin embargo, parece probable que el uso informal y exploratorio de estas técnicas siga representando una parte

importante de su uso.

Resumen

1. Las técnicas proyectivas emplean estímulos relativamente ambiguos y un formato de respuesta abierta.
2. De acuerdo con la hipótesis proyectiva, las respuestas a los estímulos ambiguos estarán determinadas primordialmente por rasgos de personalidad, motivaciones e impulsos profundamente arraigados, quizá inconscientes. La hipótesis proyectiva se asocia a menudo con un método psicoanalítico para explorar la personalidad, aunque no necesariamente ocurre esto.
3. Las técnicas proyectivas se usan mucho en la práctica clínica y también en la investigación sobre la personalidad.
4. Describir los usos de las técnicas proyectivas es complicado debido a que se aplican y califican de varios modos por distintos usuarios.
5. La Prueba Rorschach de Manchas de Tinta es la técnica proyectiva que más se usa y una de las más empleadas entre todas las pruebas psicológicas. Consta de 10 láminas, cada una con una mancha de tinta simétrica en el plano horizontal. El examinado responde a esta sencilla pregunta: ¿Qué puede ser esto?
6. Durante muchos años, hubo cinco sistemas de aplicación y calificación del Rorschach que estaban en competencia, lo que causaba confusión y resultados en conflicto. En los últimos años, el Sistema integral de Exner se convirtió en el estándar del Rorschach. El trabajo de Exner ha llevado a un resurgimiento del interés en esta prueba.
7. El *Test de Apercepción Temática* (TAT) es la segunda técnica proyectiva más usada. Consta de 30 fotos en blanco y negro (incluyendo una lámina en blanco). La aplicación clásica incluye una combinación de 20 láminas, aunque en la práctica sea común usar un número menor de láminas. El examinado cuenta una historia acerca de la imagen. Las variaciones en los procedimientos de aplicación y calificación del TAT limitan nuestra capacidad para hacer generalizaciones acerca de su confiabilidad y validez.
8. La prueba *Frases Incompletas de Rotter* (RISB) es la más usada de las numerosas pruebas de frases incompletas. Consta de 40 troncos de oraciones que el examinado debe completar de manera escrita. El RISB intenta medir la adaptación general, lo cual hace con un conjunto estandarizado de instrucciones y rúbricas de calificación.
9. Varias pruebas de dibujos de la figura humana (DFH) son herramientas clínicas muy populares. Una versión del Draw-A-Person (DAP) Test es la más usada. Ha habido muchas voces defendiendo sin sustento la validez de signos muy específicos en los dibujos. En general, las propiedades psicométricas del DFH no son alentadoras; no obstante, sirven para formular hipótesis que se pueden explorar después.
10. Las técnicas proyectivas, sin duda, seguirán siendo un artículo de primera necesidad en la práctica de los psicólogos. Sin embargo, la presión del movimiento del manejo cuidadoso puede llevar a cierta reducción de su uso. En el caso del uso formal, el campo parece moverse en dirección de procedimientos más estandarizados de

aplicación y calificación; además, hay expectativas más altas respecto de la evidencia de la confiabilidad y validez.

Palabras clave

codificación
códigos de ubicación
DAP
determinantes
DFH
estímulo ambiguo
fase de indagación
fase de respuesta
hipótesis proyectiva
HTP
KFD
populares
protocolo
Puntuaciones especiales
Resumen estructural
RISB
romper el hielo
Rorschach
Sistema integral
TAT

Ejercicios

1. Preparar sus propias manchas de tinta constituye un rito de pasaje para el psicómetra novato, pero no es tan fácil como suena. Al principio, piensas: es sólo derramar un poco de tinta en una hoja de papel y doblarla a la mitad. Sin embargo, no es así de sencillo. Para los iniciadores, era más fácil conseguir la tinta, pero en estos días –después de los bolígrafos, las máquinas de escribir y las PC– no lo es. Sugerimos usar las pinturas dactilares de los niños. El café, el refresco o líquidos similares no tienen la consistencia adecuada y sólo se absorben en el papel. Las pinturas dactilares tienen la consistencia correcta para hacer tus manchas de tinta. Asegúrate de usar pintura lavable, porque este ejercicio puede resultar muy sucio. Intenta hacer algunas con pintura negra y otras con colores pastel. Es mejor utilizar cartulina que hojas de papel. Una hoja de 21.5 × 28 cm cortada en dos, 14 × 21.5 cm, queda de un tamaño muy semejante al de las verdaderas láminas del Rorschach.
No pongas una gota de pintura en la mitad de la hoja y luego la dobles, porque eso no funciona muy bien. Antes de poner pintura en la hoja, dóblala a la mitad y luego desdóblala. Entonces pones una gota de pintura cerca del doblez, pero sólo de un lado. Luego la doblas y la presionas para que se extienda. Intenta seguir este procedimiento en varias hojas probando con diferentes números de gotas de pintura. También prueba con gotas de distintos colores. Algunas de tus producciones parecerán potencialmente útiles y otras no. Rorschach intentó con diferentes gotas, no sólo con las 10 que usamos ahora.
2. Para ilustrar los variados usos de las técnicas proyectivas, elige una base de datos electrónica como PsychINFO. Haz una búsqueda con las palabras clave "Rorschach" o "TAT". Observa la variedad de informes que resulta de esta búsqueda. ¿Puedes determinar a partir de los títulos de los informes si se ocupan primordialmente de las propiedades psicométricas de las pruebas o de su uso como variable criterio? Observa que algunos informes pueden simplemente discutir la prueba más que usarla en un proyecto de investigación.
3. Supón que deseas usar el RISB para medir *depresión*. Completa los troncos de reactivos que aparecen en el cuadro 14-5 con una respuesta que pienses que indicaría un grado *moderado* de depresión y con otra que indique una depresión *grave*.
4. Entra a este sitio de la editorial PAR (Psychological Assessment Resources): www.parinc.com. Aquí encontrarás una serie de informes del protocolo del Rorschach, incluyendo una Secuencia de puntuaciones, un Resumen estructural e informes interpretativos generados por computadora. El sitio se actualiza con frecuencia, por lo que quizá tengas que hacer una pequeña búsqueda para encontrar los informes. Por ahora, éstas son las URLS exactas, pero tienes que estar pendiente de posibles actualizaciones:

- http://www4.parinc.com/WebUploads/samplerpts/RIAP5SS.pdf

- http://www4.parinc.com/WebUploads/samplerpts/RIAP5StrucSum.pdf
- http://www4.parinc.com/WebUploads/samplerpts/RIAP5IR.pdf
- http://www4.parinc.com/WebUploads/samplerpts/RIAP5FE.pdf

¿Puedes descifrar los códigos de ubicación del Resumen estructural? La mayoría de los códigos es incomprensible para los novatos; sin embargo, los informes interpretativos deben ser comprensibles. ¿Puedes reconocer las afirmaciones con referencia a una norma que aparecen en los informes?

5. ¿Cómo diseñarías un estudio para poner a prueba la primera generalización que aparece en el cuadro 14-7?

6. Utiliza lo que sabes acerca de la relación de la media y la desviación estándar con los percentiles para estimar qué porcentaje del grupo de estandarización estaría arriba de la puntuación de corte sugerida en el RISB. La media es aproximadamente de 130 y la desviación estándar, de 17. Suponiendo que la distribución de las puntuaciones se aproxima mucho a la normalidad (lo cual así es), ¿qué porcentaje de los casos está arriba de la puntuación de corte de 145? Si es necesario, refresca tu conocimiento de estas cuestiones consultando otra vez el cuadro 3-1 y la figura 3-10a del capítulo 3.

Notas

[1] Aquí nos referimos a la cuarta edición del libro clásico de Exner: *The Rorschach*: *A Comprehensive System*: *Vol 1: Basic Foundations and Principles of Interpretation*. La primera edición, en la que se presentó por primera vez el sistema, apareció en 1974, la segunda en 1986 y la tercera en 1993. El Volumen 2 (sobre "interpretaciones avanzadas") y el Volumen 3 (sobre el uso con niños y adolescentes) aparecieron en varias ediciones de 1978 a 2005. Un tratamiento más minucioso del trabajo de Exner requeriría consultar estos volúmenes. Véase también Weiner (2003).

[2] El uso del término *primario* en estas categorías es nuestro, no de Exner. Usamos este término con fines estrictamente pedagógicos.

CAPÍTULO 15
Intereses y actitudes

Objetivos

1. Describir las principales diferencias entre las medidas de personalidad y las de intereses y actitudes.

2. Comparar los dos métodos tradicionales para la evaluación vocacional en términos del origen de las escalas y el formato de los reactivos.

3. Bosquejar el hexágono Holland y ubicar los seis códigos en sus vértices.

4. Enumerar las principales características de los inventarios de intereses vocacionales:

Strong Interest Inventory.

Kuder Career Interests Assessments.

Self-Directed Search.

5. Enumerar cinco generalizaciones principales acerca de la evaluación de intereses vocacionales.

6. Describir los métodos básicos de la medición de actitudes en los procedimientos de Likert, Thurstone y Guttman.

7. Identificar la diferencia esencial entre medición de actitudes y encuesta de la opinión pública.

Introducción

En los tres capítulos anteriores revisamos pruebas cuyo objetivo es evaluar la personalidad humana o una característica definida con exactitud, como la depresión. Para continuar explorando el dominio no cognitivo, ahora examinaremos pruebas de intereses y actitudes. La distinción entre las pruebas que tratamos en este capítulo y las de "personalidad" es muy común; sin embargo, como veremos, la distinción no es irrefutable. La diferencia más importante es el propósito de las pruebas más que su naturaleza; además, la distinción entre los términos intereses, actitudes y personalidad es borrosa, pues depende, otra vez, del propósito más que de cualquier otra cosa. Algunas pruebas de "intereses" tienen su origen en una teoría de la personalidad, pero las distinciones entre estos términos son comunes y encontraremos que aportan una manera práctica de agrupar las pruebas.

Cuando los psicólogos hacen referencia a "interés", por lo general, quieren decir intereses vocacionales o de elección de carrera (nosotros los usaremos como sinónimos). Ésta es un área amplia y aplicada de las pruebas, en especial para los orientadores. El capítulo comienza con una cobertura de este tipo de pruebas.

Pruebas de intereses vocacionales

Para hablar de las pruebas de intereses vocacionales se requiere de una orientación especial. Hay dos nombres dominantes en este campo y dos métodos primarios para elaborar pruebas con los que se asocia una serie de pruebas que, en gran parte, han definido el campo de las pruebas de intereses vocacionales. Más adelante, examinaremos las versiones más recientes de estas pruebas. Una orientación preliminar facilitará la posterior revisión de ellas.

Strong y Kuder

Los dos nombres dominantes en este campo son **Edward K. Strong, Jr.** y **G. Fredric Kuder**. Tomando en cuenta que sus nombres se encuentran con mucha frecuencia, vale la pena presentar un pequeño esbozo de ellos.

Strong empezó su trabajo sobre la medición de los intereses vocacionales en la década de 1920 mientras enseñaba en el Carnegie Institute of Technology, ahora Carnegie Mellon University, en Pittsburgh. Pronto se cambió a la Universidad de Stanford, donde realizó la mayor parte de su trabajo. La primera edición de la prueba de Strong apareció en 1927. Recordemos del capítulo 1 que en este periodo tuvieron un auge los esfuerzos por elaborar pruebas; también fue la era que vio las primeras pruebas de inteligencia de Otis, las primeras baterías estandarizadas de aprovechamiento, el Rorschach y una gran cantidad de primeras versiones en el mundo de las pruebas. En ese tiempo, Strong era casi la única persona que hacía un trabajo importante sobre la medición de intereses vocacionales. En la parte final de su vida, que terminó en 1964, Strong tuvo varios colaboradores que continuaron trabajando después de su muerte. Sus nombres a veces aparecen como coautores en la prueba de Strong, pero el nombre de Strong aparece en todas las ediciones de su prueba hasta ahora.

G. Fredric Kuder dividió sus intereses entre la medición de intereses vocacionales y la teoría de las pruebas. Recordemos las fórmulas de confiabilidad Kuder-Richardson (en especial, KR 20 y KR 21) del capítulo 4. Kuder también fue el editor fundador de *Educational and Psychological Measurement*, revista que mencionamos como publicación clave en la historia de las pruebas. El trabajo de Kuder sobre la medición de los intereses vocacionales empezó en la década de 1930 y llevó a la publicación de su primera prueba en 1939. Desde entonces, su contribución fue prolífica en este campo.

En Donnay (1997) y Zytowski (1992) se puede consultar el desarrollo histórico de los inventarios de Strong y Kuder, respectivamente. En Betsworth y Fouad (1997) se puede encontrar un resumen del campo entero de la evaluación de los intereses de elección de carrera. En *A Counselor's Guide to Career Assessment Instruments* de Whitfield, Feller y Wood (2008) se puede encontrar una lista exhaustiva de instrumentos de este campo con comentarios críticos.

Métodos tradicionales

Las medidas de intereses vocacionales difieren de dos maneras principales. La literatura de este campo está llena de referencias a estas diferencias, por lo que será útil revisarlas. La primera se relaciona con el origen de las escalas o puntuaciones de las pruebas. Un método usa una base empírica de criterio meta para elaborar las escalas; recordemos la esencia de este método que describimos en el capítulo 12. El procedimiento implica determinar qué reactivos diferencian entre grupos bien definidos; la naturaleza exacta de los reactivos y las razones para la diferenciación son inmateriales. Desde luego, el contenido de los reactivos, por lo general, será acerca de intereses y actividades relacionadas con el trabajo, pero esto no necesariamente es cierto. Para propósitos de evaluación de los intereses vocacionales, los grupos se forman a partir de las ocupaciones, por ejemplo, maestros de primaria, contadores y así sucesivamente. Obtenemos conjuntos de reactivos que diferencian cada uno de estos grupos respecto de la población general; dichos reactivos conforman una escala para cada ocupación. Desde un punto de vista práctico, con los resultados del método de criterio meta, esto es lo que un orientador le dirá a un cliente: "Sus intereses son muy similares a los de los contadores; quizá quiera considerar este campo."

La prueba original de Strong empleó el método de criterio meta. Recordemos que éste es el mismo que se usó en el MMPI, con la diferencia de que en esta prueba se trataba de grupos clínicos (deprimidos, paranoicos, etc.) y no de ocupaciones. Aunque el MMPI puede ser el ejemplo más famoso del método de criterio meta, Strong lo empleó antes que el MMPI, de modo que él es el verdadero pionero en el uso de este método en la elaboración de pruebas.

El segundo método busca producir escalas que correspondan a amplias áreas de interés, por ejemplo, artística, persuasiva o científica. Cada área puede estar relacionada con diversas ocupaciones; por ejemplo, una puntuación alta en el área de la persuasión puede sugerir un trabajo en ventas, mientras que una puntuación alta en el área artística puede sugerir artes gráficas, diseño de interiores o campos afines. Las pruebas originales de Kuder usan este método para crear las escalas.

Una segunda diferencia en el método se relaciona con el uso de *puntuaciones absolutas y relativas*. Esta diferencia surge del formato de la respuesta a los reactivos y corresponde a una interpretación absoluta o relativa de los niveles de interés. Consideremos los ejemplos del cuadro 15-1. En el formato "absoluto", al examinado le pueden gustar o disgustar todos los reactivos, mientras que en el formato "relativo", el examinado puede preferir sólo un reactivo y tiene que rechazar otro.

Cuadro 15-1. Tipos de reactivos que se usan para medir los intereses: niveles absoluto y relativo

	Nivel absoluto		
Valora el grado en que te gusta cada una de estas actividades:			
	No me gusta	Neutral	Me gusta

Disecar ranas	O	O	O
Analizar datos	O	O	O
Vender revistas	O	O	O
	Nivel relativo		
De estas tres actividades, marca con M la que más te guste y con L la que te guste menos. No marques nada en la otra actividad.			
Disecar ranas	[M]	[L]	
Analizar datos	[M]	[L]	
Vender revistas	[M]	[L]	

Resumen de puntos clave 15-1

Diferencias tradicionales en los métodos de medición de intereses vocacionales

- Origen de las escalas: Método de criterio meta o de áreas amplias
- Formato de los reactivos: Nivel de interés absoluto o relativo

Usos de las pruebas de intereses vocacionales

Los inventarios de intereses vocacionales se usan mucho, en especial en el campo de la orientación; en las encuestas a orientadores acerca de las pruebas que usan, estos inventarios, sobre todo los tres que describimos más adelante en este capítulo, por lo regular se encuentran en los primeros lugares (Bubenzer, Zimpfer, & Mahrle, 1990; Frauenhoffer *et al*., 1998; Watkins, Campbell, & McGregor, 1988; Watkins, Campbell, & Nieberding, 1994; Watkins *et al*., 1995). La mayoría de los usos tiene lugar en contextos escolares, primero en los tres años de bachillerato y luego en la universidad. También se usa un poco en los cambios de carrera que ocurren alrededor de los 40 años de edad. Incluso entre los psicólogos clínicos, ha aumentado el uso de algunos inventarios de intereses vocacionales (Camara, Nathan, & Puente, 1998).

Advertencia acerca de los nombres

El campo de las pruebas de intereses vocacionales es como una sala de espejos cuando se trata de los nombres de las pruebas, lo cual es cierto en especial tratándose de pruebas asociadas con dos nombres destacados en este campo, Strong y Kuder. A veces los nombres cambian de una edición a otra, pero a veces no. Por ejemplo, el *Strong Interest*

Inventory (**SII**) es la nueva edición del *Strong-Campbell Interest Inventory* (SCII). El SCII fue, a su vez, la nueva edición del *Strong Vocational Interest Blank* (SVIB). En ocasiones, dos ediciones diferentes del SII se distinguen sólo por la designación de las formas: T325 y T317. El SII ahora ha evolucionado en el *New Revised Strong Interest Inventory Assessment* (Donnay *et al*., 2005), que en algunos anuncios de la editorial y manuales se menciona simplemente como el *Strong Interest Inventory* o nada más como el *Strong*, lo que confirma nuestro comentario acerca de la potencial confusión entre las ediciones de las pruebas.

Del lado de Kuder, el *Kuder Preference Record–Vocational* con el tiempo se convirtió en el *Kuder Occupational Interest Survey* (KOIS), luego en el *Kuder Career Interests Assessments*. Al mismo tiempo, también hubo un *Kuder Preference Record–Personal* que buscaba ser más una medida de personalidad. También hay un *Kuder General Interest Survey*, publicado entre dos ediciones del KOIS, que usa, en parte, los reactivos del KOIS. Las referencias "al Kuder" o "al Strong" a menudo son ambiguas; incluso en la literatura profesional es común encontrar un título (o unas iniciales) cuando, en realidad, se usó una edición diferente de la prueba. En el caso de las pruebas más usadas se acostumbra citar una edición específica de la prueba, por ejemplo WAIS-IV, MMPI-2 o Stanford 10a edición, pero esta saludable práctica no se extiende a las medidas de intereses vocacionales. De hecho, a menudo parece haber un intento de ocultar las diferencias entre las ediciones. Por ejemplo, el manual del *Strong Interest Inventory* a menudo hace referencia simplemente a la investigación sobre "el Strong" sin importar si ésta se llevó a cabo con el SVIB, SCII, SII T325 o SII T317. Para ayudar al estudiante a lidiar con esta situación, en el cuadro 15-2 se presenta la genealogía de las pruebas Strong y Kuder.

Cuadro 15-2. Genealogía de las pruebas Strong y Kuder

Strong	
Strong Vocational Interest Blank (SVIB)—Men	1927
Strong Vocational Interest Blank (SVIB)—Women	1933
Strong-Campbell Interest Inventory (SCII) (formas de hombres y mujeres fusionadas)	1974, 1981
Strong Interest Inventory (SII). Forma T325	1985
Strong Interest Inventory (SII). Forma T317	1994
New Revised Strong Interest Inventory Assessment	2004
Kuder	
Kuder Preference Record—Vocational	1934
Kuder General Interest Survey (KGIS), Forma E	1963
Kuder Occupational Interest Survey (KOIS), Forma D	1966
Kuder Occupational Interest Survey (KOIS), Forma DD (mismos reactivos de la Forma D, nuevos informes de puntuaciones)	1985
Kuder Career Search with Person Match (KCS)	1999

Para complicar aún más la situación, **John Holland** tiene su propia prueba de intereses vocacionales *y* un esquema interpretativo de los intereses vocacionales (véase el hexágono de Holland en la siguiente sección). Otros autores han adoptado el esquema interpretativo de Holland. Los novatos en este campo pueden pensar que hacer referencia al trabajo de Holland es hablar de su prueba, pero con frecuencia no es así, sino que se alude al uso de su esquema interpretativo con alguna otra prueba, en especial el Strong. Todo esto es muy desconcertante para el estudiante que empieza a conocer este campo; sin embargo, esa es la situación actual.

Temas de Holland y los códigos del RIASEC

John Holland (1959, 1966, 1997) desarrolló una teoría acerca de la elección vocacional que ha brindado un método popular para informar las puntuaciones de los intereses de elección de carrera. De acuerdo con esta teoría, los intereses vocacionales se pueden organizar en seis temas o tipos principales. El cuadro 15-3 resume los seis tipos junto con algunos descriptores de personalidad y ejemplos de trabajos relacionados. Los temas son similares a los factores o dimensiones que hemos encontrado en otros contextos; sin embargo, no son igual de independientes entre sí, sino que algunos están más estrechamente relacionados que otros.

Cuadro 15-3. Tipos de personalidad de Holland con ejemplos relacionados con trabajos y características

Tipo	Código	Ejemplos de trabajos[a]	Algunos descriptores
Realista	R	guardia de seguridad, entrenador de atletismo, dentista	práctico, franco
Investigador	I	obrero siderúrgico, detective policiaco, ingeniero químico	crítico, curioso
Artístico	A	bailarín, diseñador de modas, editor	expresivo, idealista
Social	S	cuidador de niños, auxiliar de terapia ocupacional, maestro	amable, generoso
Emprendedor	E	vendedor telefónico, ventas, director de ventas	extrovertido, optimista
Convencional	C	despachador de policía, asistente dental, contador	ordenado, eficiente

[a] Para ver una lista de muchos trabajos ordenados mediante los códigos del RIASEC, entra a la página de O*NET http://www.onetonline.org/find/descriptor/browse/Interests/

Los grados de relación se pueden describir mediante un hexágono, como se muestra en la figura 15-1. Este hexágono aparece en informes de puntuaciones de varios inventarios de intereses vocacionales, así como en otros contextos. Los vértices del hexágono representan los seis temas o tipos, cuya ubicación representa el grado de relación que hay entre ellos. Los temas adyacentes tienen correlaciones moderadamente altas, mientras que los temas opuestos tienen correlaciones bajas. Podemos observar que las iniciales de los temas van en el sentido de las manecillas del reloj desde el vértice superior izquierdo: **RIASEC**. Este acrónimo señala que se usa el esquema de los tipos de Holland para

informar las puntuaciones y también da origen a un sistema de codificación que recuerda al del MMPI-2 (véase las pp. 341-342a» del capítulo 13). En particular, podemos informar los dos códigos del RIASEC más altos de un individuo; por ejemplo, el tipo de código de una persona podría ser SA (las puntuaciones más altas son Social y Artístico) o ES (las puntuaciones más altas son Emprendedor y Social). Los códigos de Holland tocaron una cuerda muy sensible de los orientadores y otros usuarios de los instrumentos de evaluación de elección de carrera. De ahí que encontramos referencias a los tipos de Holland, el hexágono y los códigos del RIASEC o simplemente al RIASEC en varios contextos dentro del mundo de la evaluación de la elección de carrera. Encontraremos los códigos del RIASEC en nuestro tratamiento de las pruebas Strong y Kuder; después, trataremos el inventario de Holland, el *Self-Directed Search*, que obviamente también emplea el esquema RIASEC.

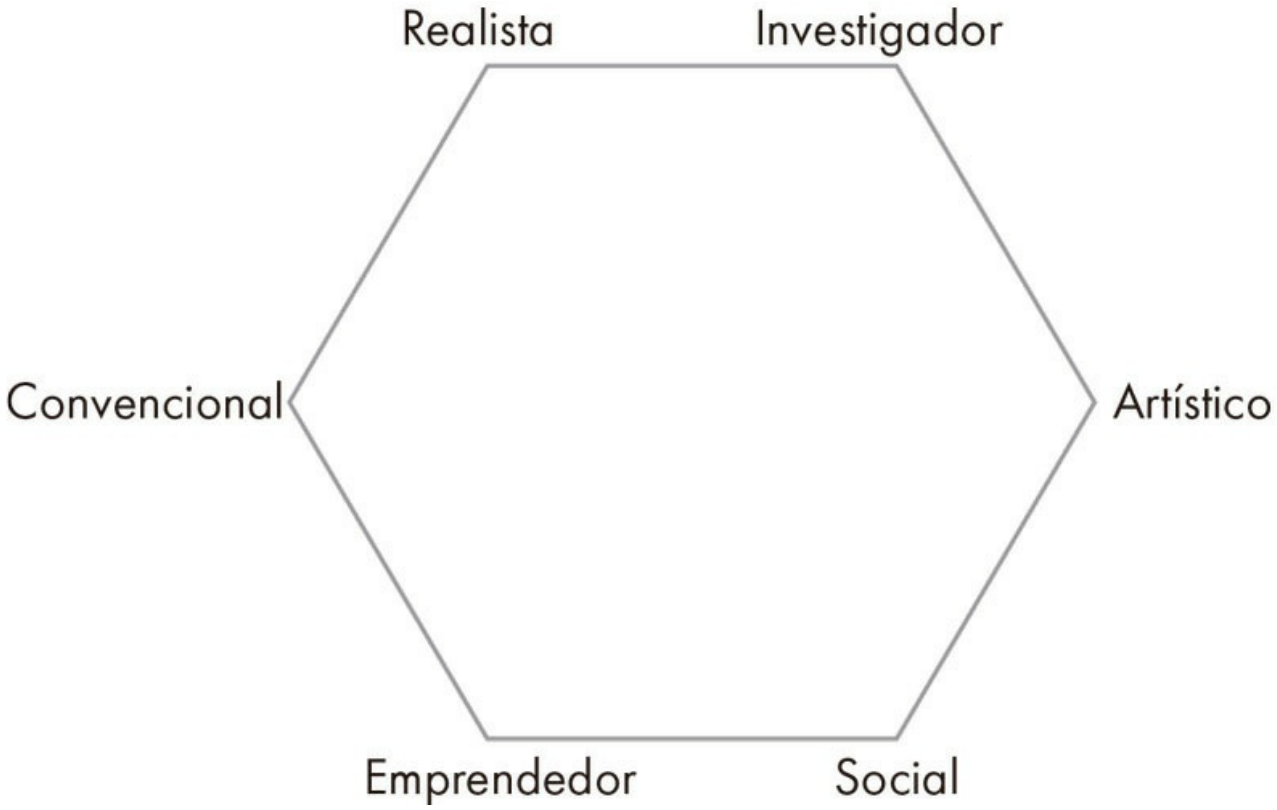
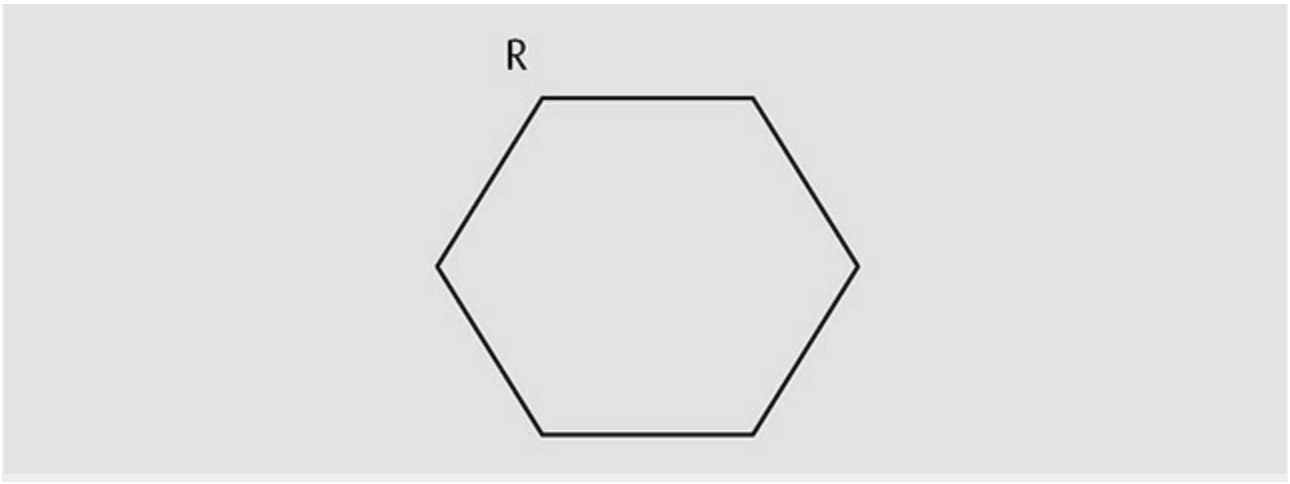


Figura 15-1. Hexágono RIASEC.

¡Inténtalo!

En las siguientes tres secciones, encontraremos con frecuencia el esquema de codificación RIASEC; memorízalo. Hazlo tanto con los nombres como con el orden correcto de los códigos. Sin ver la figura 15-1, escribe los nombres en los vértices del hexágono de Holland, empezando con R.



Resumen de puntos clave 15-2

Los tres inventarios de intereses vocacionales más importantes

1. Strong Interest Inventory
2. Kuder Career Interests Assessments
3. Self-Directed Search (SDS)

Strong Interest Inventory

Como señalamos antes, el *Strong Interest Inventory* es muy usado, por lo que es un buen candidato para hacer una presentación detallada. Ya esbozamos su historia y algunas de sus características clave, en especial el uso del método criterio meta. En las siguientes secciones, describiremos los tipos de reactivos y puntuaciones, normas, confiabilidad y validez de la nueva edición revisada del Strong.

¡Inténtalo!

Trata de pensar en otros dos reactivos para cada categoría del cuadro 15-4. Escoge uno al que responderías "Me gusta" y otro al que responderías "Me disgusta".

Cuadro 15-4. Resumen de la estructura del nuevo Strong Interest Inventory Assessment revisado

Categoría	Reactivos	Ejemplos[a]	Escala de respuesta[b]
Ocupaciones	107	Terapeuta físico	[GM] [G] [I] [D] DM]
		Conductor de *talk show*	[GM] [G] [I] [D] DM]
Áreas temáticas	46	Comunicación	[GM] [G] [I] [D] DM]
		Medicina del deporte	[GM] [G] [I] [D] DM]
Actividades	85	Analizar datos	[GM] [G] [I] [D] DM]
		Ayudar a ancianos	[GM] [G] [I] [D] DM]
Actividades de recreación	28	Jardinería	[GM] [G] [I] [D] DM]
		Viajes	[GM] [G] [I] [D] DM]
Personas	16	Líderes políticos	[GM] [G] [I] [D] DM]
		Ancianos	[GM] [G] [I] [D] DM]
Tus características[c]	9	Soy muy organizado	[MP] [P] [I] [D] [MD]
		Tengo fuertes impulsos	[MP] [P] [I] [D] [MD]
Total	291		

[a] Éstos no son reactivos verdaderos del Strong, sino ejemplos de cada categoría.
[b] GM = Me gusta mucho, G = Me gusta, I = Indiferente, D = Me disgusta, DM = Me disgusta mucho.
[c] La escala de respuesta cambia en Tus características, y va de Muy parecido a mí a Muy distinto a mí.

Estructura y tipos de reactivos

El nuevo Strong (Donnay *et al.*, 2005) contiene 291 reactivos agrupados en seis categorías, como se describe en el cuadro 15-4. En general, el examinado responde una serie de descriptores sencillos (p. ej., jardinería, análisis de datos) marcando el grado en que le gusta cada uno. En esta versión se usa una escala de respuesta de cinco puntos,

mientras que en todas las versiones anteriores la escala era de tres puntos (Me gusta, Indiferente, Me disgusta). Las instrucciones dicen: "No dediques mucho tiempo a pensar en cada uno [de los reactivos]. Básate en tu primera impresión". La aplicación del SII toma de 35 a 40 min. Está dirigido a estudiantes de bachillerato y universidad, así como a adultos.

Tipos de puntaciones

El SII produce cinco tipos de puntuaciones, que se enumeran en el cuadro 15-5. El número total de puntuaciones separadas es de más de 200, así que es útil observarlas por tipo. Estudia este cuadro con cuidado, pues en las secciones posteriores haremos referencia a estos tipos de puntuación con frecuencia.

Cuadro 15-5. Tipos de puntuaciones producidas por el Strong Interest Inventory

1. Temas Ocupacionales Generales (TOG)	6 puntuaciones
Se basan en el hexágono de Holland. Las puntuaciones son Realista, Investigador, Artístico, Social, Emprendedor y Convencional (RIASEC).	
2. Escalas de Intereses Básicos (EIB)	30 puntuaciones
Se basan en el análisis factorial de los reactivos. Las puntuaciones son de áreas como Atletismo, Ciencia, Milicia, Ventas.	
3. Escalas Ocupacionales (EO)	260 ocupaciones
Las escalas clásicas Strong emplean el método de clave del criterio con varios grupos ocupacionales.	
4. Escalas de Estilo Personal	5 puntuaciones
Aparecieron en la edición de 1994. Las puntuaciones son de estilo de Trabajo, ambiente de Aprendizaje, estilo de Liderazgo, toma de Riesgos, orientación al Equipo.	
5. Índices de Aplicación	3 resúmenes
Total de respuestas. Índice de Tipicidad. Porcentajes dentro de las categorías de respuesta.	

Los Temas Ocupacionales Generales (TOG) corresponden a las categorías de Holland, que describimos antes (figura 15-1). Representan una amplia área de intereses. Las Escalas de Intereses Básicos (EIB) también representan las áreas de intereses generales, pero tienen una definición más restringida. Para propósitos del informe, las EIB se agrupan dentro de los TOG; por ejemplo, las EIB de derecho, administración y ventas se agrupan en el TOG Emprendedor.

Las Escalas Ocupacionales (EO) son escalas tradicionales de varios grupos ocupacionales hechas con el método de criterio meta. Una puntuación alta en una de estas escalas, digamos asistente de abogado o trabajador social, significa que el examinado tiene intereses similares a los de los individuos que ya se desempeñan en esa ocupación. En las ediciones previas, no todas las ocupaciones tenían normas para hombres y mujeres, pero la edición actual proporciona normas por género para todas las ocupaciones. A partir de 2012, la lista de las EO se actualizó: se agregaron algunas

escalas nuevas, se revisaron otras y se eliminaron otras más (Herk & Thompson, 2012). La actualización de 2012 no afectó los otros tipos de puntuaciones enumeradas en el cuadro 15-5.

Las Escalas de Estilos Personales, que aparecieron en el SII de 1994, comprenden una extraña mezcla de cinco puntuaciones relacionadas con las preferencias ambientales y las maneras de relacionarse con las personas y situaciones. Los Índices de Aplicación ofrecen información acerca de la sinceridad con que el examinado respondió el inventario. En el lenguaje del capítulo 12 sobre las pruebas objetivas de personalidad, se trata de los índices de validez. El Índice Total de Respuestas muestra el número de reactivos marcados (de los 291): los informes no se generan si hay menos de 276. El Índice de Tipicidad se basa en 24 pares de reactivos en que se esperarían respuestas similares. Se registra una advertencia si hay inconsistencias en 17 o más de estos pares. El informe de los Porcentajes de Respuestas a los Reactivos simplemente muestra la distribución de las respuestas en los cinco puntos de la escala. Porcentajes excepcionalmente grandes en alguna categoría sugieren que se debe tener cuidado al interpretar ese protocolo; por ejemplo, ¿cómo interpretarías el hecho de que un examinado respondió "Indiferente" en 75% de los reactivos?

Normas

La descripción de los grupos de estandarización del Strong requiere de una explicación cuidadosa (véase aclaración adicional en la figura 15-2). Primero, hay una Muestra Representativa General (MRG) conformada por 2250 casos, divididos equitativamente por género. La MRG sirvió como base para desarrollar un sistema de puntuación estándar con $M = 50$ y $DE = 10$; se trata del conocido sistema de puntuaciones T, muy usado con otras pruebas. Estas puntuaciones estándar se aplican a las escalas TOG y EIB.

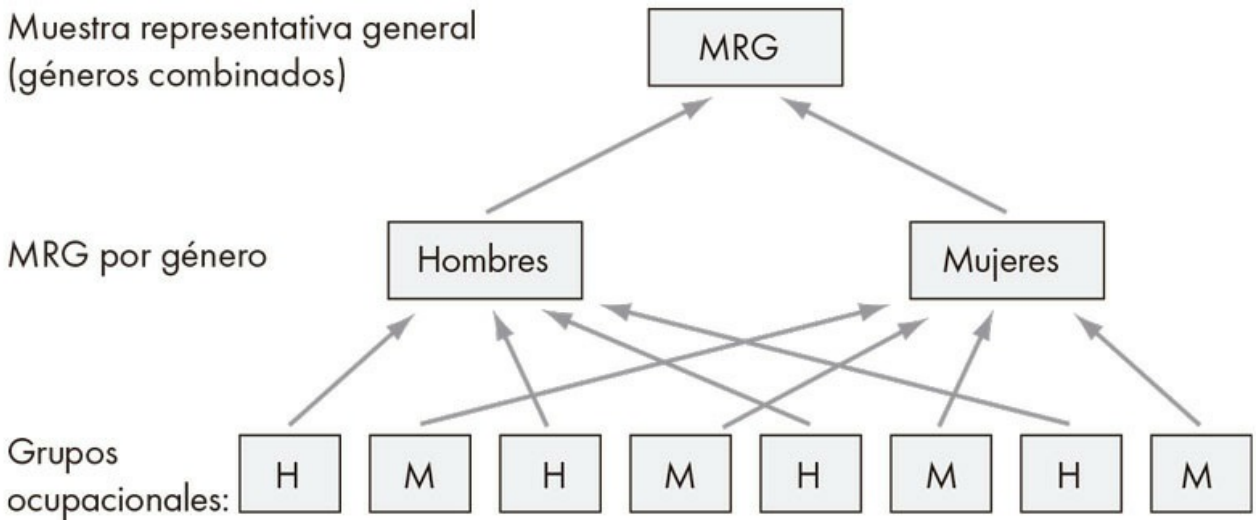


Figura 15-2. Esquema de los grupos de estandarización del Strong.

Sin embargo, se desarrollaron sistemas de puntuaciones estándar separados (pero también con M = 50 y DE = 10) para las Escalas Ocupacionales, cuyas normas evidentemente derivan de muestras de cada ocupación. Estas muestras van más allá de los casos de la Muestra Representativa General que ya hemos descrito. Se usó una gran variedad de métodos para reclutar a los participantes de estos grupos de normalización. En general, las personas de cada uno de estos grupos tenían que a) realizar los deberes principales del puesto, b) tener al menos tres años de experiencia en el puesto y c) expresar satisfacción con el puesto. El manual del Strong admite con franqueza la dificultad para asegurar muestras adecuadas y el sesgo potencial introducido por la naturaleza por completo voluntaria (y un poco pesada) de la participación. Los grupos ocupacionales de normalización varían de manera considerable en tamaño y años en que se recabaron los datos.

Así, hay varios sistemas de puntuación estándar separados en operación. Además, todas las escalas de percentiles (del TOG, EIB y EO) están separadas por género; por ejemplo, una puntuación estándar de 50 en el TOG Realista muestra que la persona se encuentra justo en el promedio de la población general, mientras que una puntuación estándar de 50 en la EO Contador muestra que una persona se encuentra justo en el promedio de los contadores, que son muy diferentes de la población general en varios sentidos. En el caso de una mujer, la puntuación de 50 la ubicará en un percentil (en las normas de las mujeres) mucho más alto que si se tratara de un hombre (y utilizando las normas de los hombres). La figura 15-3 bosqueja la estructura del informe de las puntuaciones del *Strong Interest Inventory*.

Temas Ocupacionales Generales (GOT) – En realidad, las 6 áreas RIASEC
Áreas RIASEC ordenadas por rango.
Las tres áreas más altas producen un código de temas (p. ej., CEA).

Escalas de Intereses Básicos – 30 áreas
Ordenadas por rango, con énfasis en las 5 áreas más altas.

Escalas Ocupacionales – 122 ocupaciones
Ordenadas por rango, con énfasis en las 10 ocupaciones más altas.

Escalas de Estilo Personal
Gráfica sencilla de percentiles de las 5 escalas.

Índices de Aplicación
Resumen sencillo; no hay comparaciones normativas.

Nota: Se proporcionan puntuaciones estándar de todas las escalas. Los percentiles de los rangos siempre son por género.

Figura 15-3. Bosquejo de la estructura de un informe del Strong Interest Inventory.

En conjunto, una gran variedad de diferentes estructuras normativas constituye la base de las puntuaciones del Strong, lo cual subraya la necesidad de dominio técnico de estos temas al interpretar los resultados. Esta advertencia se aplica a casi todos los inventarios de intereses vocacionales, y no sólo al Strong.

¡Inténtalo!

Para ver un ejemplo de un informe real de 14 páginas del Strong, entra a https://www.cpp.com/sampleReports/reports.aspx y haz clic en uno de los perfiles del Strong.

Confiabilidad

El manual del Strong presenta extensos datos de confiabilidad de todas las escalas, a excepción de los Índices de Aplicación. Se informa la consistencia interna (alpha de Cronbach) de los TOG y las EIB con base en la Muestra Representativa General. El alpha media de los TOG es .92, y en ninguna escala es menor de .90. El alpha media de las 30 EIB es .87, y ninguna es menor de .80. En el caso de las escalas de Estilo Personal, el coeficiente medio de alpha es .86. De manera extraña, los coeficientes alpha

de las EO no se informan.

La confiabilidad de test-retest de diferentes grupos se informa en el caso de todas las puntuaciones, a excepción, otra vez, de los Índices de Aplicación. Casi todas las confiabilidades de test-retest de los TOG se ubican alrededor de .85, y ninguna es menor de .80. También la mayoría de estas confiabilidades de test-retest de las EIB se encuentran alrededor de .85, y ninguna es menor de .75. En el caso de las escalas de Estilo Personal, las confiabilidades de test-retest estuvieron, por lo general, alrededor de .85. Es curioso que no se proporcionen datos de confiabilidad de los Índices de Aplicación.

¿Qué concluimos acerca de esta información? Primero, es evidente que los intereses vocacionales son notablemente estables, lo cual es confortante para los orientadores y otros profesionales que necesitan basarse en esta información. Segundo, en los pocos casos en que los datos de confiabilidad son apenas aceptables, por ejemplo, cuando son menores de .75, el problema, casi de manera invariable, es la brevedad de la escala. Recordemos el principio general de que la confiabilidad depende en gran medida del número de reactivos de la escala. Algunas de las del SII, en especial las EIB, tienen sólo 5 o 6 reactivos nada más. Tercero, observamos con disgusto la ausencia de datos de confiabilidad de los Índices de Aplicación, pues son necesarios para interpretar cualquier información, aunque esté expresada de manera distinta a las puntuaciones convencionales.

Validez

¿Cómo se demuestra la validez de una medida de intereses vocacionales? ¿Existe un criterio que sea predicho? ¿Existe un rasgo conceptual que se deba documentar? Estas preguntas no admiten respuestas sencillas. Hay dos métodos comunes para demostrar la validez de este tipo de medidas y una gran cantidad de métodos auxiliares. El primer método es mostrar que los resultados de la prueba diferencian entre grupos ocupacionales existentes en direcciones predecibles. Por ejemplo, si hay una escala "maestro" en el inventario, las personas que practican esa profesión deben tener puntuaciones altas en esa escala, y deben ser más altas que en las demás escalas. El segundo método para demostrar la validez de una medida de intereses vocacionales es mostrar que las puntuaciones predicen la ocupación (o carrera universitaria) que las personas eligen en última instancia. Por ejemplo, debe suceder que las personas con una puntuación alta en la escala "maestro" tiendan a convertirse en maestros más que en, digamos, vendedores. A la inversa, las personas con puntuaciones altas en la escala "ventas" tienden a convertirse en vendedores más que, digamos, en científicos investigadores.

El manual del Strong contiene una gran cantidad de estudios a lo largo de las líneas que hemos descrito. De hecho, entre los inventarios de intereses vocacionales más usados, el Strong muestra con claridad la información más amplia de validez. Teniendo en cuenta la importancia histórica del método de criterio meta para las escalas ocupacionales de esta prueba, es de especial relevancia la cuestión del grado en que estas escalas diferencian

personas de una ocupación respecto de la población general, por lo que es comprensible que el manual del Strong dedique atención considerable a esta comparación. Para propósitos de análisis estadístico, esta cuestión se convierte en un tema de tamaño del efecto en una comparación de grupos: justo la situación que describimos en el método de contraste grupal al discutir los procedimientos de validación en el capítulo 5. En la figura 15-4, reproducimos la figura pertinente de ese capítulo renombrando los grupos de contraste. La *d* de Cohen = $(M_1 - M_2) / DE$ describe el grado de diferenciación.[1] Mediante este método, la *d* mediana de las EO es 1.51, que se encuentra en un punto intermedio entre los dos ejemplos de la figura 15-4. El estadístico *d* varía de 1.00 a 2.64, valores respetables que, sin embargo, forman un rango considerable.

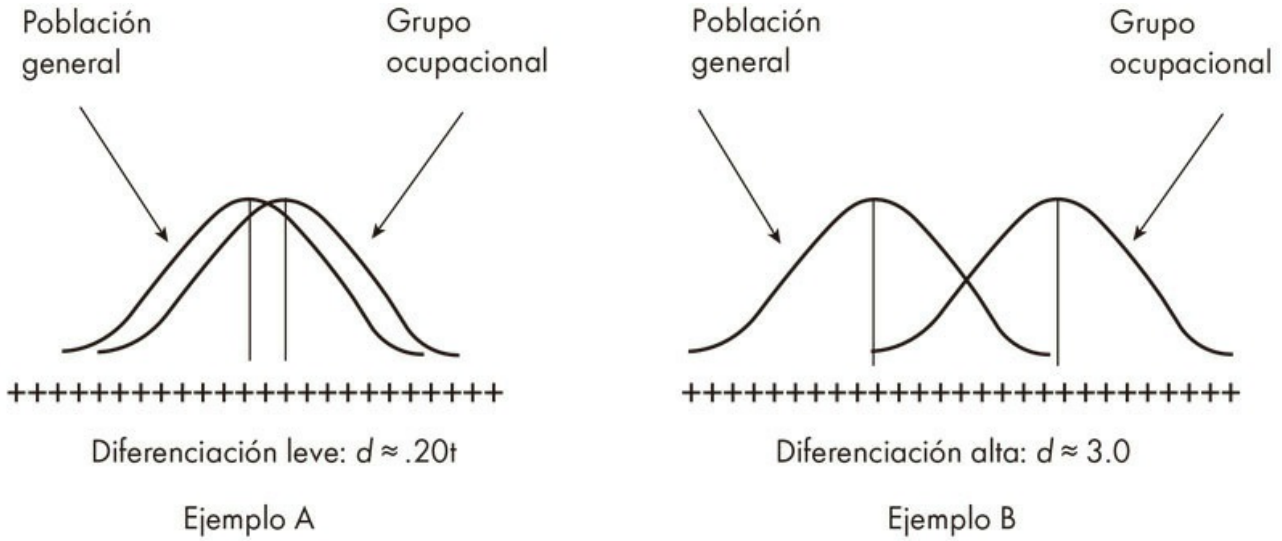


Figura 15-4. Grados de diferenciación de las Escalas Ocupacionales del Strong.

El manual del Strong presenta numerosos estudios que muestran que las personas con puntuaciones altas en una escala particular tienden, en última instancia, a elegir ocupaciones consistentes con esa puntuación. En un método usado en varios estudios se dio seguimiento de 3 a 18 años a personas con puntuaciones de 45 o más en una escala para determinar sus ocupaciones. Casi en la mitad de los casos, las personas tenían en efecto un empleo consistente con sus puntuaciones en el Strong; desde luego, esto significa que cerca de la mitad no tenía un empleo de esta naturaleza. El manual también contiene información de peso acerca de las correlaciones entre las escalas que, en general, apoya la validez de constructo de la prueba.

Un problema importante con la interpretación de gran parte de la información sobre la validez es que muchos estudios se llevaron a cabo con versiones anteriores de la prueba. El manual del Strong afirma continuamente que estos estudios deberían ser aplicables a la versión actual; sin embargo, es casi un acto de fe aceptar estas afirmaciones.

Kuder Career Interests Assessments

El *Kuder Career Interests Assessments* (**KCIA**), una trilogía de inventarios, es el último en la larga línea de instrumentos Kuder (véase la genealogía en el cuadro 15-2).[2] Reemplaza varias formas anteriores del Kuder y emplea muchos de los conceptos de sus antecesores. El precursor inmediato del KCIA fue el *Kuder Career Search with Person Match* (KCS; Zytowski, 2005). La editorial comercializa el KCIA como parte del Sistema de Planeación de Carrera Kuder, que incluye un portafolio de carreras electrónico, un sistema de administración de datos y varios elementos más.

Estructura y reactivos

El KCIA emplea tríadas de elección forzada, método distintivo del Kuder que describimos antes en este capítulo. En cada tríada, el examinado marca la actividad que más le gusta, la que le sigue y la que le menos le gusta. (En ediciones anteriores, sólo se marcaba la actividad que le gustaba más y la que gustaba menos, y la tercera se dejaba en blanco). En la figura 15-5 se ofrece una "foto" de lo que ve la persona que contesta el inventario. Ahí podemos notar las tríadas de reactivos.

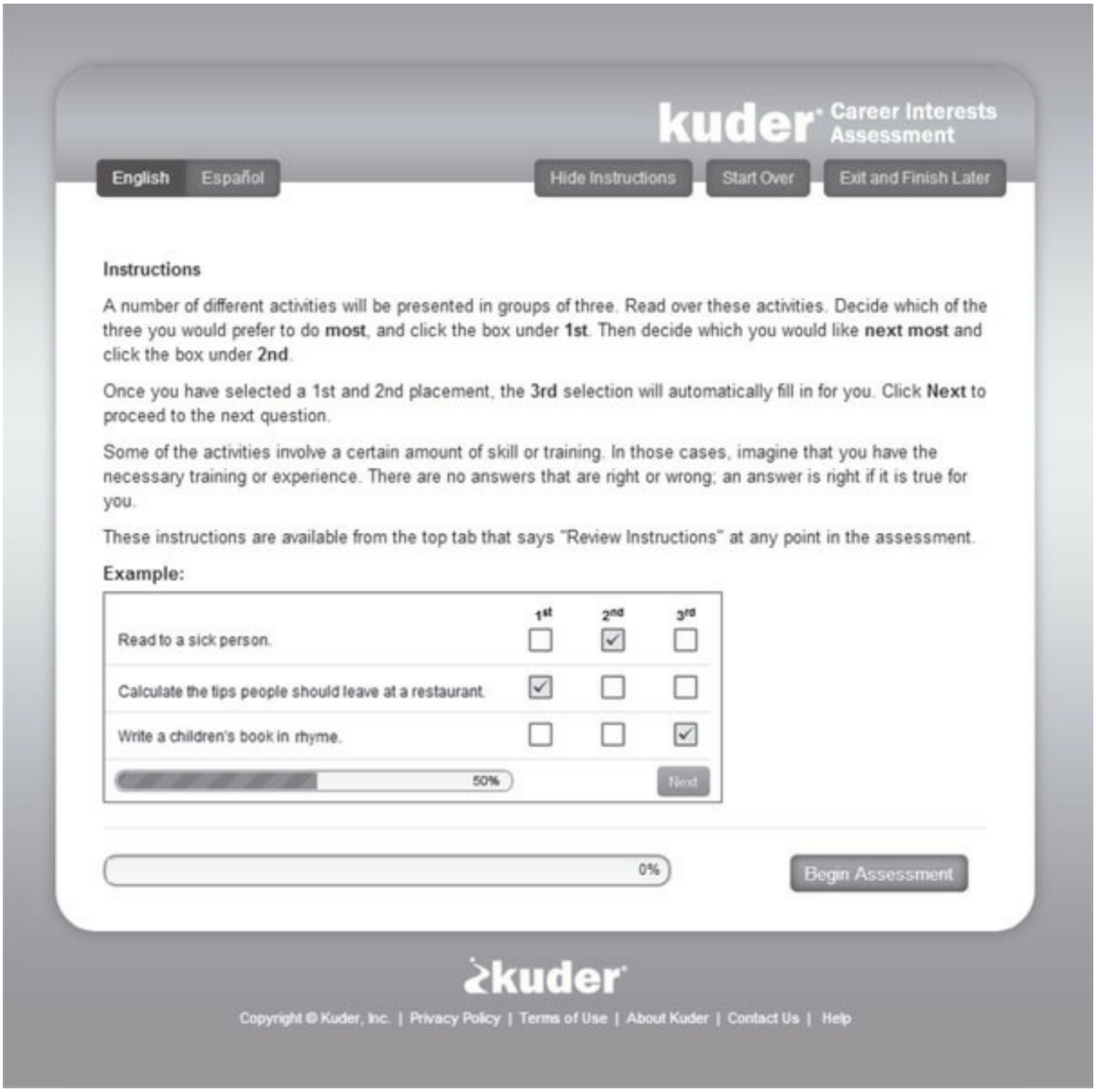


Figura 15-5. Imagen de las instrucciones para contestar el *Kuder Career Interests Assessment*.
Fuente: Copyright © Kuder, Inc., www.kuder.com. Reproducido con autorización

¡Inténtalo!

Para familiarizarte con el formato de elección forzada del Kuder, marca tus respuestas a los reactivos de la figura 15-5.

El KCIA ahora está disponible en tres niveles: *Galaxy* para estudiantes de primaria, *Navigator* para estudiantes de secundaria y bachillerato y *Journey* para estudiantes universitarios y adultos.

Puntuaciones

El KCIA emplea varias puntuaciones, entre las que se incluye un método innovador de informar: el *person match*. Un tipo de puntuaciones usa un conjunto de 16 "grupos" de ocupaciones, cuyo propósito es ser escalas relativamente homogéneas que representen dimensiones importantes de los intereses, de manera muy similar a las escalas del *Strong Basic Interest*. La figura 15-6 muestra parte de un reporte basado en el nivel *Journey* del KCIA. Podemos notar que en el informe aparece un perfil de grupos ordenados de acuerdo con su rango; en el caso de la persona de la figura 15-6, el interés más alto es Servicios humanos y el más bajo, Ciencia, tecnología, ingeniería y matemáticas.

Figura 15-6. Imagen de los resultados de grupos nacionales del Kuder Career Interests Assessment.
Fuente: Copyright © Kuder, Inc., www.kuder.com. Reproducido con autorización

Otro método para informar implica emparejar al examinado con otras personas que ya están en el mundo del trabajo y tienen perfiles similares del Kuder. (En la figura 15-6, observa el botón de emparejamiento a la derecha del informe.) Los perfiles se construyen con base en las puntuaciones de los grupos. El examinado obtiene una lista de dichas personas junto con descripciones de ellas y sus actividades. Al hacer clic en una persona (digamos, emparejamiento 2), se muestra una descripción del trabajo de esa persona basada en cuestionarios contestados por 2000 personas (fondo de emparejamientos), cuyas respuestas se editan para representar una minientrevista. Esta metodología da un giro interesante a la continua búsqueda de los psicólogos de mejores maneras de interpretar la información de las pruebas; el tiempo dirá qué tanto éxito logra.

¡Inténtalo!

Para ver muestras de los innovadores informes de emparejamiento, entra en http://www.personmatch.kuder.com/pmsamples.aspx

Confiabilidad y validez

A partir de 2013, no hay un manual técnico del KCIA disponible: una sorprendentemente larga demora después del lanzamiento de las nuevas pruebas en 2009. Hay información del KCS disponible (Zytowski, 2005), pero los datos de confiabilidad son escasos. Por ejemplo, las confiabilidades KR-20 que se informan corresponden a una muestra de sólo 146 adultos de quienes no se presenta más información. Estas confiabilidades varían de .70 a .80, con una mediana de .73: baja de manera desconcertante. El manual del KCS también informa correlaciones de puntuaciones de Grupos de Carreras con las del Strong Interest Inventory y las del Self-Directed Search (que describiremos a continuación) basadas sólo en 77 universitarios de quieres no se presenta más información. Las correlaciones varían de .12 a .57, con una mediana de .50: de nuevo desconcertantemente baja.

Self-Directed Search (SDS)

El *Self-Directed Search* (**SDS**) se hace publicidad como el inventario de intereses vocacionales más usado. Sea el primero, segundo o tercero en rango, el SDS es con claridad un instrumento muy popular, lo cual es algo sorpresivo, porque no cuenta con la distinguida historia, los informes elaborados y otras características de los inventarios Strong y Kuder. De hecho, cierta informalidad y su poca cultura parecen ser parte de su atractivo. El manual del SDS hace hincapié en que el propio examinado se aplica el inventario, lo califica y lo interpreta. El SDS busca la sencillez, casi como si renunciara a toda sofisticación técnica, a puntuaciones múltiples y al procesamiento por computadora de los inventarios Strong y Kuder. Esta sencillez parece explicar su relativamente rápido crecimiento en popularidad.

El SDS apareció por primera vez en 1971, y en 1977, 1985 y 1994 se publicaron nuevas ediciones. Hay varias formas de la edición actual, diseñada principalmente para diferentes grupos de edad. Aquí nos concentraremos en la Forma R de la edición de 1994, dirigida a estudiantes de bachillerato, universidad y adultos. Ésta es la forma que más se usa y con la que más investigaciones se han hecho (desde 2012, un SDS revisado se encuentra en la fase de estandarización).

El SDS tiene su origen en el esquema RIASEC de John Holland (Holland, 1959, 1966, 1997). Puedes revisar la figura 15-1 para ver los códigos de este esquema. De acuerdo con la teoría de Holland, existen seis tipos básicos de personalidad, así como seis tipos de ambientes de trabajo. Cuando hay una buena correspondencia entre el tipo de personalidad y el del ambiente de trabajo, el resultado debe ser satisfacción con la carrera. Recordemos que los inventarios Strong y Kuder incorporan los códigos RIASEC, mientras que el SDS, en su totalidad, se basa en dicho sistema.

Estructura y reactivos

El SDS consta de 228 reactivos y su aplicación requiere cerca de 30 min. Los reactivos aparecen en cuatro secciones importantes; dentro de cada una, los reactivos se agrupan de acuerdo con el código RIASEC. El cuadro 15-6 muestra estas secciones, el número de reactivos en cada una, el formato de respuesta y la naturaleza de los reactivos. La calificación implica nada más sumar el número de respuestas Sí y Me gusta, y las autoevaluaciones dentro de cada área a lo largo de las cuatro secciones. Esto produce seis puntuaciones naturales a las que se denomina Puntuaciones escalares de resumen, una por cada área RIASEC. La puntuación máxima por área es 50 (11 “Me gusta” en Actividades en un área, 11 “Sí” en Competencias, 14 “Sí” en Ocupaciones y dos “7” en Autoevaluaciones), mientras que la puntuación mínima es cero. Después de sumar las puntuaciones, la persona determina las tres puntuaciones naturales más altas, lo cual produce un código RIASEC de tres letras.

La quinta sección del SDS, Sueños diurnos ocupacionales, no se califica de manera

formal, pero puede ser útil al interpretar los resultados. En esta sección se pide a la persona que "enumere las carreras con las que ha soñado despierto, así como las que ha discutido con otros". La persona también encuentra los códigos RIASEC que corresponden a estas ocupaciones.

El *Explorador de Carreras y Ocupaciones* es un complemento esencial del resumen de las puntuaciones del SDS –el código RIASEC de tres letras. Se trata de un breve cuadernillo que enumera cientos de títulos ocupacionales de acuerdo con el código RIASEC; en un principio, estos códigos implicaban títulos del *Dictionary of Occupational Titles* (DOT; U.S. Department of Labor, 1991), pero éste se convirtió en 2010 en el sistema de clasificación de trabajos O*NET (véase esa sección más adelante).

Cuadro 15-6. Organización del Self-Directed Search (SDS)

Sección	**Número de reactivos por área RIASEC**	**Respuesta**	**Naturaleza de los reactivos**
Actividades	11	Me gusta/Me disgusta	Actividades que podrían gustarte hacer
Competencias	11	Sí/No	Cosas que sabes hacer bien
Ocupaciones	14	Sí/No	Ocupaciones que te interesan
Autoevaluaciones	2	Escala de 7 puntos	Autovaloración de la capacidad
		(7 = alto, 1 = bajo)	

Normas

La interpretación de las puntuaciones del SDS depende, en su mayor parte, de las Escalas de puntuaciones de resumen, que son naturales, traducidas a códigos RIASEC. Así, el SDS ofrece un método de referencia a un criterio para la interpretación, aunque en los procedimientos de selección de reactivos se incorporaron elementos del método normativo. Las normas no son parte importante del esquema interpretativo; no obstante, el manual del SDS ofrece normas de rangos de percentiles para estudiantes de bachillerato, de universidad y muestras de adultos por separado, que después se dividen por género dentro de cada grupo. El cuadro 15-7 resume los tamaños de las muestras de cada grupo. El manual presenta poca información sobre los procedimientos para seleccionar a los individuos de los grupos de estandarización; de ahí que lo mejor es verlos como muestras por conveniencia con todas las limitaciones que eso implica.

Cuadro 15-7. Número de casos por grupo para las normas del SDS

	Grupo de edad		
Género	**Bachillerato**	**Universidad**	**Adultos**
Hombres	344	399	251

Mujeres	475	713	405

Confiabilidad y validez

El manual del SDS presenta las confiabilidades de consistencia interna (KR-20), basadas en las muestras de estandarización que hemos descrito, de las escalas de Actividades, Competencias y Ocupaciones, las cuales son de poco interés, ya que parece que éstas nunca se interpretan con independencia respecto de las Puntuaciones escalares de resumen. También se presentan las confiabilidades de consistencia interna de estas puntuaciones; todas se encuentran en el rango de .90 a .94. Las confiabilidades de test-retest, basadas en muestras muy limitadas (n = 73) e intervalos de 4 a 12 semanas entre las aplicaciones, variaron de .76 a .89. Es claro que se necesita más información acerca de la estabilidad temporal del SDS.

El manual del SDS aborda la validez desde diferentes ángulos. Como es común en los inventarios de intereses vocacionales, la información sobre la validez del SDS tiende a ser una amalgama de datos de las sucesivas ediciones de la prueba. Primero, intenta mostrar que las escalas son razonablemente independientes una de otra y consistentes con el modelo hexagonal de Holland (figura 15-1). Segundo, como ocurre con otros inventarios de este tipo, hay numerosos estudios sobre "tasa de aciertos" que muestran el grado en que las puntuaciones del SDS son consistentes con criterios como carrera universitaria u ocupación real. Por último, hay muchos estudios sobre la correlación entre el SDS y una gran cantidad de pruebas de personalidad como el NEO PI (véase pp. 323-326a»). El interés en las relaciones con este tipo de pruebas es un poco inusual tratándose de un inventario de elección de carrera; sin embargo, es consistente con la filosofía del SDS, según la cual la elección de carrera es una cuestión de emparejar el tipo de personalidad con el ambiente de trabajo.

El manual del SDS contiene un capítulo innovador sobre efectos y resultados, que en la terminología que introdujimos antes corresponde a la validez consecuencial, pero en este manual no se emplea dicho término. El capítulo presenta estudios de si el SDS hace alguna diferencia en la vida de las personas. El manual (Holland, Fritzsche, & Powell, 1997) señala que "a menos que esta experiencia [contestar, calificar e interpretar el SDS] haga una diferencia, los altos niveles de confiabilidad y validez no tienen ningún uso práctico" (p. 57). Por ejemplo, ¿contestar el SDS lleva a tener mayor claridad acerca de la elección vocacional? ¿Lleva a considerar más opciones de carrera? ¿Cómo se compara contestar esta prueba y tener una sesión con un orientador vocacional? Los resultados de estos estudios sobre efectos y resultados son, por lo general, aunque no siempre, favorables para el SDS. Más importante para los propósitos de este libro, el capítulo del SDS señala una dirección saludable para otros manuales de pruebas.

¡Inténtalo!

Sin contestar en verdad el SDS, trata de ubicarte en los códigos RIASEC (1 es alto y 6, bajo). "Ordena" las áreas, no las "valores".

Área	Realista	Investigador	Artístico	Social	Emprendedor	Convencional
Rango	_____	_____	_____	_____	_____	_____

Con base en este orden, ¿cuál es tu código de tres letras? _____

*O*NET* «394a

Todos los inventarios de intereses vocacionales tienen referencias de O*NET, un proyecto del U.S. Department of Labor [Secretaría del trabajo]. O*NET enumera las ocupaciones y muestra los niveles de educación necesarios e información relacionada. Es una fuente invaluable de información para orientadores y estudiantes que exploran las opciones de carrera. Es interesante que O*NET use el sistema RIASEC de Holland para clasificar las ocupaciones por intereses.

¡Inténtalo!

Puedes contestar tu propio inventario de intereses basado en el RIASEC de manera gratuita en http://www.mynextmove.org/explore/ip. Si esta liga no funciona, sólo escribe "O*NET Interest Profiler" en cualquier buscador de internet. El Interest Profiler consta de 60 reactivos y toma cerca de 3 min contestarlo. Obtendrás un informe basado en el sistema RIASEC que señala tus posibles ocupaciones.

Algunas generalizaciones acerca de las medidas de intereses vocacionales

¿Qué generalizaciones podemos hacer sobre las medidas de intereses vocacionales, al menos con base en el examen de las tres que más se usan? Podemos sugerir varias observaciones.

1. Los patrones de los intereses vocacionales parecen ser bastante confiables, al menos de la adolescencia media en adelante de acuerdo con los resultados de estas pruebas. Esto es cierto en términos de la consistencia interna y estabilidad temporal. Sin embargo, las confiabilidades no son perfectas y, en algunos casos, son bastante bajas. Los informes formales de las puntuaciones tienden a mencionar estos hechos relacionados con la confiabilidad menos que en otras áreas del campo de la evaluación. Por ejemplo, aunque en los manuales técnicos se informa el error estándar de la medición de las puntuaciones, esta información no suele incorporarse en los informes.

2. En particular, si empleamos en última instancia la elección de carrera como criterio, parece que las medidas de intereses vocacionales tienen un grado de validez respetable. Los grupos ocupacionales tienden a diferir en sus patrones de intereses, y las personas tienden a adoptar ocupaciones consistentes con sus intereses. Sin embargo, debemos hacer una advertencia en este aspecto. Todos los manuales de las pruebas y otros informes de investigación, de manera comprensible, tienden a hacer hincapié en las diferencias promedio entre los grupos ocupacionales mientras ocultan la superposición en las distribuciones. Además, los informes destacan que la correspondencia entre las puntuaciones de las pruebas y las entradas ocupacionales está por encima de lo que se esperaría por simple azar, al mismo tiempo que no prestan atención al considerable número de casos cuando hay una falta de correspondencia.

3. Las medidas de intereses vocacionales carecen notablemente de referencias a técnicas psicométricas más modernas. Incluso en los manuales más recientes de los inventarios más usados hay pocas o ninguna referencia a la teoría de la respuesta al reactivo, funcionamiento diferencial de reactivos, coeficientes de generalizabilidad, validez consecuencial, etc. Para estar seguros, algunos de estos conceptos son muy pertinentes para los inventarios de intereses vocacionales, y los manuales los tocan sólo de manera tangencial. Sin embargo, el uso formal de las técnicas y la terminología de la psicometría moderna parecen no haber llegado al campo de la evaluación de los intereses vocacionales, o viceversa.

4. Cada vez más, los inventarios de intereses vocacionales se contestan en línea y el usuario tiene de inmediato los resultados. Las editoriales de los tres inventarios que revisamos en este capítulo promueven de manera categórica este modo de aplicación y de elaboración de informes, aunque aumentan la carga para ellas. La aplicación tradicional era, por lo común, en un salón de clases, y los informes se enviaban a un orientador para que los revisara junto con los estudiantes (u otros clientes) de manera

individual o grupal. Al menos, los usuarios sabían con quién discutirían los resultados. Esto no sucede con la aplicación y elaboración de informes en línea, por lo que éstos deben tratarse con precaución debido a su imperfecta confiabilidad y validez, si bien estas precauciones no siempre son evidentes en los informes.

5. En los inicios de su trabajo, Strong supuso que había una relación positiva entre intereses y capacidad para luchar por ellos, pero pronto encontró que esta relación no existía, lo cual se confirmó con claridad después. Una persona puede querer ser artista, pero no tener la habilidad para tener éxito en su intento por conseguirlo; otra persona puede querer ser físico, pero no tener la capacidad para terminar los cursos de ciencia necesarios. Por ello, desde el punto de vista práctico, la evaluación de los intereses vocacionales debe complementarse con información pertinente del dominio de las capacidades. En la actualidad, el inventario de intereses está acompañado de una guía (o una sección dentro del inventario) relacionada con las capacidades. Sin embargo, estas guías o secciones a veces se basan por completo en autoevaluaciones de las capacidades; por ejemplo, en una escala de 1 a 7, ¿qué tan bueno eres en matemáticas? Se debe prestar mucha atención a la relación entre estas autoevaluaciones y las medidas reales de las capacidades. No siempre es claro que esto se haga. Otra alternativa es obtener puntuaciones de una medida de capacidad cognitiva, como las que describimos en el capítulo 9, o de aprovechamiento, como las que revisamos en el capítulo 11.

Resumen de puntos clave 15-3

Generalizaciones acerca de las medidas de intereses vocacionales

- Bastante confiables
- Grado de validez respetable
- Escaso uso de las técnicas psicométricas modernas
- Mayor uso de la aplicación en línea
- Evaluación de las capacidades además de los intereses

Medidas de actitudes

Los psicólogos adoran las medidas de actitudes, lo cual es comprensible porque nuestras actitudes hacia un sinfín de temas dan colorido a nuestras vidas: actitudes hacia el trabajo, hacia otras personas, hacia los partidos y temas políticos, entre muchas otras. Todas son características humanas importantes. La medición de actitudes tiene un papel destacado en el campo de la psicología social. Aquí hay algunos ejemplos:

- ¿Cuáles son las actitudes de los niños hacia los inmigrantes?
- ¿La actitud hacia la donación de órganos se relaciona con los registros reales para la donación?
- ¿Cómo cambia la actitud hacia la religión conforme avanza la edad?
- ¿La actitud hacia los ancianos es un rasgo unidimensional?

Primero debemos indagar qué distingue las actitudes de los intereses o de los rasgos de personalidad. Es común usar estos términos como si fueran distintos, pero sus límites son borrosos.

De hecho, a veces los puntos de referencia de estos términos no pueden diferenciarse; por ejemplo, la extroversión en ocasiones se puede pensar como un rasgo de personalidad y en otras como una actitud. Los pensamientos y sentimientos de una persona acerca de un puesto de trabajo pueden verse como un interés vocacional o como una actitud. No obstante, la medición de actitudes suele tratarse por separado respecto de la medición de los rasgos de personalidad y los intereses vocacionales.

Una actitud suele definirse respecto de su objeto, que puede ser un concepto (p. ej., democracia), una práctica (p. ej., la pena capital), un grupo (p. ej., italoamericanos), una institución (p. ej., la universidad) o un individuo (p. ej., Dr. Norcross). Las posibilidades son casi ilimitadas: puede haber una actitud hacia cualquier cosa. La mayoría de los investigadores piensan que las actitudes tienen tres componentes. El primero es el cognitivo: pensamientos sobre el objeto, en especial conscientes y articulados. El segundo componente es el emocional: sentimientos hacia el objeto. Y el tercero es el conductual: las acciones que se realizan, o pueden realizarse, respecto del objeto.

La mayoría de nuestras medidas de actitudes es de lápiz y papel (o sus clones en una pantalla de computadora). Una persona responde preguntas acerca de los objetos o reacciona frente a afirmaciones acerca del objeto. Estas preguntas o afirmaciones son los troncos de los reactivos; por lo común, las respuestas son de opción múltiple, por ejemplo, De acuerdo-En desacuerdo, Sí-No, etc. Tomando en cuenta este formato, concluimos que la mayoría de las medidas de actitud abordan el componente cognitivo. Aunque algunos reactivos pueden preguntar por los sentimientos (componente emocional) o las acciones (componente conductual), las respuestas informan lo que la persona piensa acerca de sus sentimientos y conductas, pero no son una medida directa de estos componentes.

Si bien la mayoría de las mediciones de actitudes implica informes de lápiz y papel, hay otras técnicas. Las medidas fisiológicas (p. ej., respuesta galvánica de la piel o dilatación de la pupila) pueden ser de especial utilidad para investigar el componente emocional, mientras que mediante la observación se puede obtener información del componente conductual. En esta sección, nos concentraremos en las medidas de lápiz y papel porque son las que más se usan.

Existe una cantidad prácticamente infinita de actitudes, por lo que esto mismo se puede decir de las medidas de actitudes. Shaw y Wright (1967), por ejemplo, presentan cientos de medidas específicas de actitudes. Dos de los capítulos más extensos del *Directory of Unpublished Experimental Mental Measures* [Directorio de Medidas Mentales Experimentales No Publicadas] (Goldman & Mitchell, 2008) contienen medidas de "actitudes" y "valores". Robinson, Shaver y Wrightsman (1991) también ofrecen numerosos ejemplos de estas medidas. Una búsqueda reciente en ETS *Test Collection* arrojó no menos de 2641 entradas para "attitude" [actitud]. Una regla general es: si puedes pensar en una actitud, casi con total certeza, alguien ha creado una escala para medirla. Aunque hay muchas medidas de actitudes, ninguna se usa tanto como se usan las pruebas MMPI, SAT o WISC. Además, la mayoría de estas medidas tiene bases de investigación escasas en comparación con las pruebas más usadas. En el resto de los tipos de pruebas que describimos en este libro, identificamos algunas que se usan mucho y luego las describimos; sin embargo, este método no funciona en el área de las medidas de actitud.

Lo que distingue de manera especial a las medidas de actitud es el método de construcción de escalas. En la literatura profesional, a menudo hay referencias a uno de estos métodos cuando se describe una medida específica, por lo que incumbe al estudiante conocer estos métodos y sus implicaciones acerca de la naturaleza de la escala. Existen numerosos métodos, pero sólo describiremos los tres que más se citan (véase Resumen de puntos clave). El lector puede consultar las siguientes referencias clásicas para ver las descripciones de otros métodos, así como un tratamiento más detallado de los que revisaremos aquí: Edwards (1957), Shaw y Wright (1967) y Torgerson (1958).

Resumen de puntos clave 15-4

Métodos más usados para construir escalas de actitud

- Método Likert de evaluaciones sumarias
- Método de Thurstone de intervalos aparentemente iguales
- Análisis de escalograma de Guttman

Escalas Likert

Por mucho, el método de Likert (1932) es el que más se usa en la actualidad para construir escalas de actitud. En la literatura técnica a menudo se denomina **método de evaluaciones sumarias**. En su forma más pura, el método **Likert** tiene las siguientes características. Primero, empieza con una gran cantidad de troncos de reactivo que expresan algún aspecto de la actitud hacia cierto objeto. La noción básica es que cada reactivo da cierta información acerca de la actitud de una persona. La suma de toda esta información define una actitud holística hacia el objeto. La figura 15-7 ayuda a ilustrar el método Likert.

Segundo, presenta respuestas en la siguiente escala de cinco puntos: Totalmente de acuerdo, De acuerdo, No sé o Neutral, En desacuerdo, Totalmente en desacuerdo. En el cuadro 15-8 se puede ver un conjunto muestra de respuestas. Las respuestas con etiquetas verbales suelen convertirse en un número; los esquemas para asignar los valores numéricos a estas respuestas varían. Una práctica común es asignarles valores de 5, 4, 3, 2 y 1, respectivamente. Otra alternativa es usar +2, +1, 0, −1 y −2, respectivamente. Tercero, se realiza un análisis de reactivos para seleccionar los que tienen correlaciones más altas con la puntuación total, la cual es la suma de valores numéricos de las respuestas a los reactivos individuales: de ahí el nombre de método de evaluaciones sumarias. Por último, se construye una escala final con base en los resultados del análisis de reactivos.

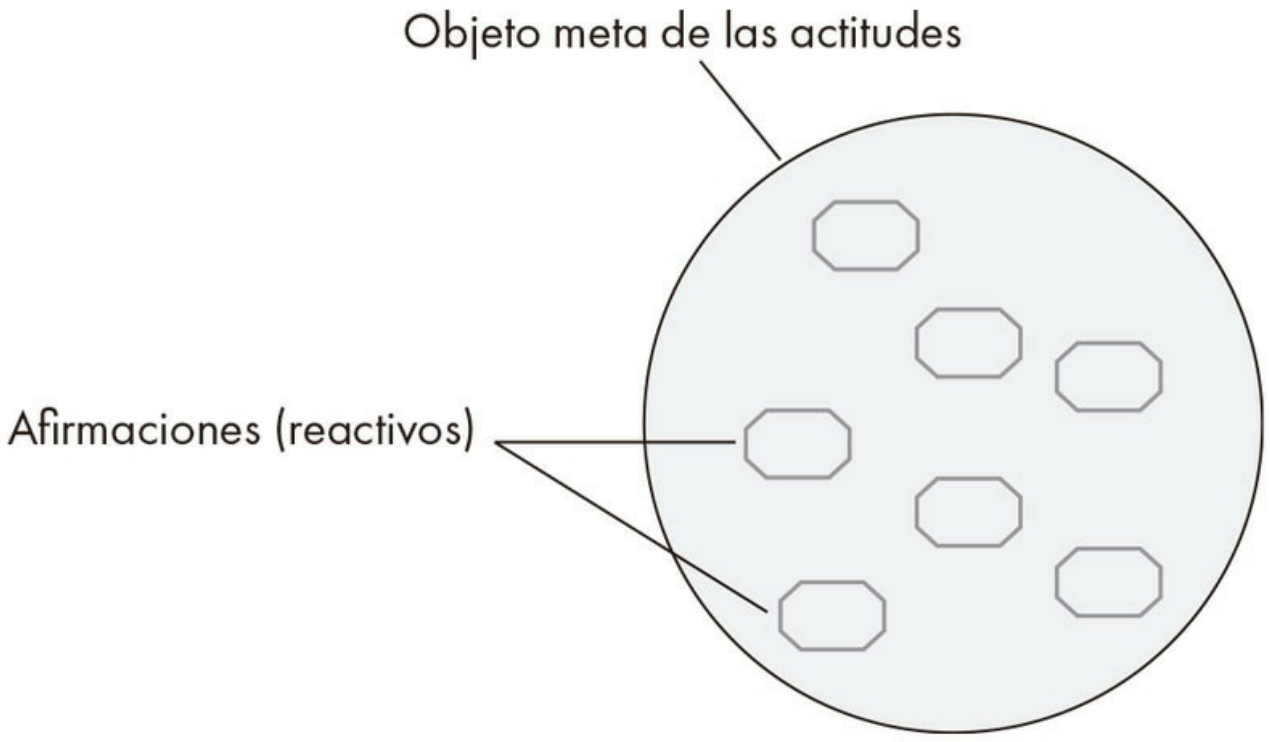


Figura 15-7. Ilustración esquemática del método Likert para la medición de actitudes.

Cuadro 15-8. Respuestas muestra de una escala Likert de 15 reactivos

Reactivo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Tom	TA	A	A	TA	TA	N	A	TA	A	TA	TA	A	N	A	TA
Dick	A	N	N	D	D	A	A	N	A	D	D	N	D	N	A
Harry	D	TD	TD	TD	TD	N	N	D	D	TD	TD	D	D	TD	D
Tom tiene una actitud favorable hacia el tema: la mayoría de sus respuestas son A y TA.															
Dick tiene una actitud neutral o sentimientos encontrados: una mezcla de respuestas N, D y A.															
Harry tiene una actitud desfavorable hacia el tema: la mayoría de sus respuestas son D y TD															

¡Inténtalo!

Asigna estos valores a las respuestas del cuadro 15-8 en el caso de Dick:
TA = +2 A = +1 N = 0 D = −1 TD = −2
¿Cuál es la puntuación de Dick en los 15 reactivos?

Hay diversas variaciones comunes en el último método de construcción de escalas Likert que describimos. Primero, en términos técnicos, la metodología Likert exige una escala de respuesta de cinco puntos; sin embargo, es común encontrar variaciones en este aspecto. Las personas usan cualquier número de categorías de respuesta: 3, 7, 10, 25, 99, y aún así lo llaman método Likert. Muchos estudios han señalado las ventajas relativas de usar distintos números de puntos en la escala de respuesta (véase, p. ej., Cheng, 1994; Johnson & Dixon, 1984; Matell & Jacoby, 1972). Segundo, las etiquetas de las categorías de respuesta varían de manera considerable; como señalamos antes, el método original de Likert empleaba las categorías Totalmente de acuerdo–Totalmente en desacuerdo. Sin embargo, también se emplean con frecuencia otras etiquetas, por ejemplo (citando sólo los puntos extremos), Total aprobación–Total desaprobación, Me gusta mucho–Me disgusta mucho, Casi siempre–Casi nunca. El autor de la escala puede adaptar las categorías a los troncos de los reactivos; de hecho, en algunos casos, las etiquetas de las respuestas pueden variar de un grupo de reactivos a otro dentro de la misma escala de actitud. Por ejemplo, los primeros cinco reactivos pueden emplear el formato Totalmente de acuerdo–Totalmente en desacuerdo, mientras que los siguientes cinco pueden usar el formato Total aprobación–Total desaprobación.

Tercero, en la forma simple del método Likert, todos los reactivos tienen la misma **direccionalidad**; por ejemplo, las respuestas Totalmente de acuerdo siempre indican una actitud favorable hacia el objeto. Sin embargo, en la práctica es común invertir la direccionalidad de algunos reactivos, de modo que un extremo del continuo de respuestas a veces corresponde a una actitud favorable y otras, a una actitud desfavorable. Esta variación ayuda a controlar la posible influencia de la dirección de las respuestas, decir siempre sí o no. Recordemos que discutimos esta dirección de las respuestas en las páginas 315-317a». Podemos notar la diferencia en la direccionalidad en los reactivos de la figura 15-8.

	TA	A	N	D	TD
1. El gobierno trabaja con bastante eficiencia.	0	0	0	0	0
2. La mayoría de los empleados de gobierno tiene un salario excesivo.	0	0	0	0	0
3. Los programas gubernamentales son esenciales para la sociedad.	0	0	0	0	0
4. La iniciativa privada debe reemplazar la acción del gobierno.	0	0	0	0	0

Nota 1: TA = Totalmente de acuerdo, A = De acuerdo, N = Neutral, D = En desacuerdo, TD = Totalmente en desacuerdo.
Nota 2: Los reactivos 2 y 4 deben invertirse para propósitos de la calificación

Figura 15-8. Reactivos muestra de una escala tipo Likert para medir la actitud hacia el gobierno.

Por último, el método Likert supone que una dimensión actitudinal subyace en los reactivos. En la práctica, los autores a menudo analizan factorialmente las respuestas a los reactivos para determinar si más de una dimensión subyace en los reactivos. Por ejemplo, el análisis factorial de los reactivos relacionados con actitudes hacia las matemáticas puede revelar que hay dos dimensiones relativamente independientes: actitud hacia la utilidad de las matemáticas y actitud hacia resolver problemas matemáticos. Así, la escala de actitudes producirá dos puntuaciones. Desde luego, en este escenario, la puntuación total que se utiliza para el análisis de reactivos se basa en los reactivos dentro de un único factor.

El formato Likert de las escalas de actitudes se usa mucho en la actualidad. En el caso de muchas actitudes es fácil construir reactivos; el formato de respuesta es bastante flexible. Y la metodología de la investigación para seleccionar reactivos y determinar la confiabilidad de consistencia interna es bastante sencilla. Por lo tanto, encontramos escalas Likert en una gran cantidad de aplicaciones.

Escalas Thurstone

La segunda metodología popular para construir escalas de actitudes es el método de Thurstone. En realidad, Thurstone elaboró varios métodos para medir actitudes; véase una lista práctica en Thurstone (1959) y Edwards (1957). Sin embargo, cuando hablamos del método Thurstone para medir actitudes, por lo general nos referimos al método de **intervalos aparentemente iguales** (Thurstone & Chave, 1929). La mayoría de las aplicaciones contemporáneas del método de Thurstone no incluye todos sus intentos por crear una escala sofisticada en términos psicofísicos. Éstos son los elementos de la metodología que se emplean en la actualidad. Primero, escribe una gran cantidad de afirmaciones que expresen actitudes hacia el grupo meta; las afirmaciones deben cubrir todos los posibles matices de la opinión (véase figura 15-9). Piensa en opiniones que

vayan desde las más favorables hasta las más desfavorables, del amor al odio. El formato de respuesta para todas las afirmaciones es De acuerdo–En desacuerdo; en la práctica, se puede pedir al examinado que simplemente marque las afirmaciones con las que está de acuerdo. En la figura 15-10 se puede ver un ejemplo. Segundo, pide a un grupo de jueces que clasifique las afirmaciones en 11 categorías, de las más favorables a las menos favorables.

Marca A (De acuerdo) o D (En desacuerdo) por cada reactivo.

Valor de la escala			
1.2	Estas personas son el peor grupo del mundo.	A	D
1.4	Este grupo casi no causa problemas.	A	D
5.2	Como casi todos los grupos, éste es una mezcla de buenos y malos.	A	D
5.4	Tengo sentimientos encontrados hacia este grupo.	A	D
10.2	Estas personas hacen una contribución muy positiva al mundo	A	D
10.8	Este grupo es lo máximo.	A	D

Figura 15-9. Ejemplos de reactivos de una escala tipo Thurstone para medir las actitudes hacia un grupo.

Se supone que las categorías son equidistantes una de otra, de modo que la diferencia entre las categorías 5 y 6 debe ser la misma que entre las categorías 8 y 9: de ahí el nombre de intervalos aparentemente iguales. Thurstone estaba muy preocupado por estas distancias psicológicas; sin embargo, podemos aplicar la metodología sin obsesionarnos con este punto. Tercero, determina para cada afirmación la ubicación promedio de la categoría y alguna medida de variación en ésta; por ejemplo, determina la media y la desviación estándar de las ubicaciones de las categorías en el caso de la afirmación número 23. Cuarto, elimina afirmaciones que tengan desviaciones estándar grandes, ya que en ellas los jueces no tuvieron un acuerdo suficiente. Quinto, agrupa las afirmaciones con valores similares de las categorías, por ejemplo, todas las afirmaciones con valores promedio de 6.0-6.9. Por último, elige algunas afirmaciones que representen posiciones actitudinales a lo largo del continuo, desde las menos hasta las más favorables. Si empiezas con 100 o 150 afirmaciones, la escala final podría tener 30. En la figura 15-9, aparecen afirmaciones muestra con sus valores asociados de la escala. Observa la diferencia en “lo favorable” expresado por estas afirmaciones; luego observa cómo estos valores de la escala funcionan en el conjunto de reactivos de la figura 15-10.

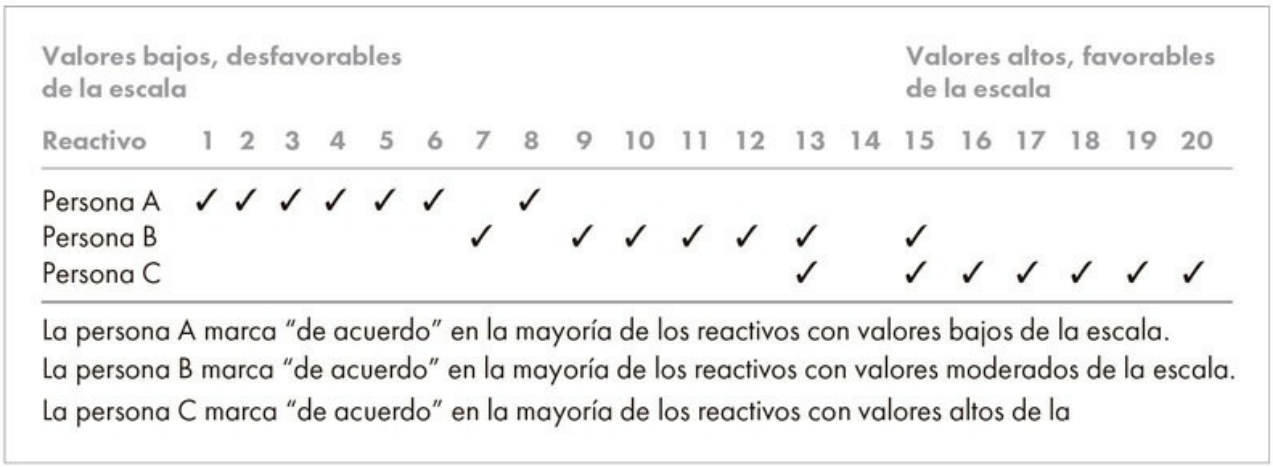


Figura 15-10. Ilustración esquemática de cómo se ubican las actitudes de tres personas en una escala Thurstone.

¡Inténtalo!

Regresa a la figura 15-9. Escribe un reactivo que pienses que pueda tener un valor en la escala de 7.0. Desde luego, en la práctica, tu reactivo habría sido valorado por un grupo de jueces.

Después del uso inicial de Thurstone de esta metodología, muchos otros autores desarrollaron escalas tipo Thurstone; sin embargo, la metodología ya no se usa tanto en la actualidad. Parece que se requiere demasiado esfuerzo para llevar a cabo la investigación necesaria para el escalamiento Thurstone, sobre todo en comparación con la metodología Likert. No obstante, las nociones básicas que desarrolló Thurstone han tenido una gran influencia en la teoría de la respuesta al reactivo.

Escalas Guttman

El tercer método para construir escalas de actitudes que se cita con frecuencia en la literatura profesional es el de **Guttman**. Su nombre técnico es **análisis de escalograma**; sin embargo, suele denominarse sólo escalamiento Guttman en honor de su autor (Guttman, 1944, 1947; también véase Guttman & Suchman, 1947).[3] Aunque este método se cita con frecuencia, en realidad no se aplica en muchos casos, porque sus requerimientos técnicos son demasiado estrictos. Supone mayor consistencia, incluso rigidez, en las actitudes de las personas de la que en realidad existe. De ahí que las ilustraciones del escalamiento Guttman a menudo empleen ejemplos triviales, incluso tontos. No obstante, este escalamiento es una herramienta conceptual útil, no sólo para la medición de las actitudes, sino también para la medición de otros rasgos como inteligencia, aprovechamiento o depresión.

La idea básica del escalamiento Guttman es obtener un conjunto de reactivos ordenados con una consistencia interna total. Entonces, ubicamos la posición de una persona dentro de este conjunto ordenado de reactivos; si la escala cumple con los criterios de Guttman, sabemos de qué modo la persona responderá *todos* los reactivos. El truco es encontrar el punto de cambio de la persona en los reactivos, pues éste describe por completo el patrón de respuestas de la persona. Ésta responde todos los reactivos de un lado del punto de cambio en una dirección y todos los del otro lado en la dirección opuesta. Consideremos las respuestas de la figura 15-11; hay ocho reactivos de actitudes a los que las personas A, B y C responden mostrando su acuerdo (+) o desacuerdo (−) con cada uno de ellos. El patrón de respuestas de las tres personas es consistente con lo que hemos dicho acerca del punto de cambio.

Reactivos	1	2	3	4	5	6	7	8
Persona A	+	+	–	–	–	–	–	–
Persona B	+	+	+	+	–	–	–	–
Persona C	+	+	+	+	+	+	–	–

Figura 15-11. Ilustración de respuestas en una escala Guttman.

Para crear una escala Guttman (es decir, para hacer un análisis de escalograma), determinamos si las respuestas se ajustan a un patrón como los de la figura 15-11. Si hay muchas inconsistencias (p. ej., + + − + − +), significa que los reactivos no forman una escala Guttman. Un término importante en el escalamiento Guttman es **replicabilidad**, a veces llamada sólo "rep", que nos dice el grado en que el patrón de respuestas *es* por completo consistente. A quienes elaboran escalas Guttman les gusta ver una replicabilidad de al menos 90%.

En nuestro ejemplo, aplicamos la noción del escalamiento Guttman a la medición de actitudes. Las respuestas + y – corresponden a "de acuerdo" y "en desacuerdo"; sin embargo, también pueden corresponder a "correcto" e "incorrecto" en una prueba de aptitud matemática. Esto ilustra cómo el concepto de escalamiento Guttman puede aplicarse a otras áreas. La idea básica se relaciona estrechamente con nociones de la teoría de la respuesta al reactivo (TRR).

Sondeos de opinión pública y estudios de mercado

En algún punto, la medición de las actitudes se convierte en sondeo de opinión pública. ¿Cuál es la diferencia esencial? Sin duda, *no es la naturaleza de las preguntas*. Los reactivos de las escalas de actitudes pueden, sin ningún problema, aparecer en un sondeo

de opinión pública y viceversa. La diferencia esencial es el objetivo de las inferencias que se hacen a partir de los resultados. En el caso de las medidas de actitud, el objetivo primordial es un individuo; queremos ubicarlo en un continuo actitudinal. En el caso de los sondeos de opinión pública, el objetivo es un grupo; lo que nos interesa en primer lugar es la posición de un grupo de personas. Por ejemplo, podemos estar interesados en determinar el porcentaje de personas que votarán por cierto candidato, qué porcentaje está a favor de la pena capital o qué marca de cereal prefiere la gente. Hablar de marcas de cereal nos recuerda que los sondeos de opinión pública incluyen el campo de los estudios de mercado, donde nuestro interés se centra en las actitudes de las personas hacia algún producto o servicio. Al igual que con otros sondeos de opinión pública, los estudios de mercado se centran en resultados grupales.

Resumen

1. Las diferencias entre las pruebas de personalidad, que vimos en capítulos anteriores, y las de intereses y actitudes dependen más del propósito de la evaluación que de la naturaleza de las pruebas mismas.
2. Los inventarios de intereses vocacionales se usan mucho en psicología, en especial en el campo de la orientación vocacional.
3. En la actualidad, hay tres inventarios predominantes: *Strong Interest Inventory*, *Kuder Career Interests Assessments* y *Self-Directed Search*.
4. Los métodos tradicionales de la medición de intereses vocacionales se dividen en dos dimensiones: el origen de las escalas (criterio meta o áreas amplias) y el formato de los reactivos (absoluto o relativo).
5. Los nombres de Edward K. Strong Jr. y Fredric Kuder han dominado este campo a lo largo de ediciones sucesivas de sus inventarios por más de 50 años. En años más recientes, el trabajo de John Holland ha sido destacado.
6. El hexágono de Holland, el esquema de codificación RIASEC, ofrece un sistema para hacer informes que se usa con mucha frecuencia, así como una base teórica para el instrumento Self-Directed Search.
7. El *New Revised Strong Interest Inventory* (SII) tiene 291 reactivos en seis bloques. Produce más de 200 puntuaciones en cinco categorías, que se comunican en un informe con un formato muy colorido. El SII cuenta con amplia información sobre su confiabilidad y validez, obtenida en muchos estudios durante un largo periodo.
8. El *Kuder Career Interests Assessments*, el último en la línea de los inventarios de intereses vocacionales Kuder, emplea reactivos de elección forzada y produce puntuaciones en 16 áreas de interés vocacional. Su innovador método para elaborar informes empareja al examinado con personas que ya tienen un empleo en ocupaciones pertinentes.
9. En busca de la sencillez, el Self-Directed Search (SDS) destaca que es el propio examinado quien se lo aplica, califica e interpreta. Los códigos RIASEC de Holland ofrecen la base teórica de la estructura del SDS y el formato del informe. Los códigos están ligados a ocupaciones organizadas en el sistema RIASEC.
10. Hacemos cinco generalizaciones acerca de los inventarios de intereses vocacionales más usados: comentarios sobre su confiabilidad y validez, falta relativa de técnicas psicométricas modernas, uso de la aplicación en línea y la importancia de medir capacidades relacionadas con los intereses.
11. Hay miles de medidas de actitudes específicas y ninguna se usa mucho más que otras. Describimos tres métodos para construir escalas de actitud: los métodos Likert, Thurstone y Guttman.
12. Los sondeos de opinión pública usan preguntas similares a las de las medidas de actitud, pero su foco de atención son los resultados grupales y no los del individuo.

Palabras clave

análisis de escalograma
direccionalidad
Guttman
Holland, John
intervalos aparentemente iguales
KCIA
Kuder, G. Fredric
Likert
método de evaluaciones sumarias
replicabilidad
RIASEC
SDS
SII
Strong, Edward K. Jr.
Thurstone

Ejercicios

1. Observa el informe del Kuder en la figura 15-6. Supón que se trata del perfil de un alumno de último año del bachillerato que trata de decidir una carrera universitaria. ¿Qué consejo podrías darle a este estudiante? ¿Qué información, además de los resultados del Kuder, debería considerar el estudiante antes de elegir su carrera?
2. Examina el conjunto de respuestas que aparecen en el cuadro 15-8 a reactivos Likert. Aplica lo que aprendiste en el capítulo 4 acerca de la confiabilidad de consistencia interna. ¿Las respuestas muestran un alto grado de consistencia interna? Explícalo brevemente.
3. Inventarios de intereses vocacionales como el Strong y el Kuder, por lo general, proporcionan datos normativos basados en grupos ocupacionales y carreras universitarias. Algunos de estos datos pueden ser de hace 40 años. ¿Crees que los patrones de intereses de los grupos ocupacionales y carreras universitarias siguen siendo, en esencia, los mismos? Trata de pensar en dos ocupaciones y dos carreras universitarias que podrían seguir siendo las mismas y otras dos que hayan cambiado de manera considerable. ¿Cómo estudiarías la estabilidad de los patrones de intereses con el paso del tiempo?
4. Aunque los procedimientos para realizar un análisis (de escalograma) Guttman de una medida de actitud pueden ser abrumadores, el SPSS te permite realizarlo con facilidad. En el SPSS, haz clic en Analysis, luego en Scale y en Reliability Analysis. En la pestaña de Model, haz clic en Guttman. Inventa un pequeño conjunto de datos como los de la figura 15-11, pero usa valores numéricos en vez de + y −. Aplica el procedimiento Guttman y observa el resultado.
5. En la sección de sondeos de opinión pública, señalamos que la naturaleza de las preguntas (reactivos) de estos sondeos a menudo es la misma que en las medidas de actitudes. Para observar esta semejanza, visita la página de internet de agencias de sondeos nacionales. Haz una búsqueda en internet, por ejemplo, de sondeos Gallup o Roper. Revisa algunas preguntas de cualquier sondeo reciente. ¿Estas preguntas podrían usarse en una medida de actitudes de lápiz y papel?
6. Para observar la enorme variedad de actitudes que los psicólogos han investigado, haz una búsqueda en una base de datos como PsychINFO con la palabra clave "attitude". Para no tener una lista demasiado extensa, limita la búsqueda a los últimos dos años. Sólo observa la diversidad de puntos de referencia de "attitude". También puedes tener un buen panorama general de la gran variedad de medidas de actitudes si consultas algunas de las fuentes citadas en el capítulo 2.
7. Imagina que estás creando el *Strong Interest Inventory*. Añade dos reactivos que pienses que pueden ser de utilidad a cada parte, I-III, que se muestran en el cuadro 15-4.

Parte	Nuevos reactivos

I. Ocupaciones	_____	_____
II. Asignaturas escolares	_____	_____
III. Actividades	_____	_____

8. ¿En qué características especiales de los inventarios de intereses vocacionales hacen hincapié las editoriales? Visita estos sitios de internet y descúbrelo:

- Strong: www.cpp.com/strong/
- Kuder: www.kuder.com
- SDS: www.self-directed-search.com o www.parinc.com

9. Varios sitios te permiten contestar un inventario de intereses vocacionales basado en el sistema RIASEC. Aquí hay dos en los que lo puedes intentar. Sin embargo, debes tener en cuenta que estos inventarios no tienen la información técnica que está disponible en el caso de instrumentos como el Strong, Kuder y SDS.
http://personality-testing.info/tests/RIASEC.php
http://www.mynextmove.org/explore/ip

Notas

1 El manual del Strong emplea el índice Q de Tilton, que casi siempre tiene una correlación perfecta con la *d* de Cohen.

2 Para ver una historia en imágenes de la serie Kuder de 1938 a 2013, entra a http://www.75yearsofkuder.com/timeline/

3 Guttman (1947, p. 247) dijo: "Lo llamaremos técnica Cornell para el análisis de escalogramas". Pero este nombre no tuvo una buena aceptación.

CAPÍTULO 16
Aspectos éticos y legales

Objetivos

1. Distinguir entre consideraciones éticas y legales.
2. Citar las principales fuentes de los principios éticos para el uso de pruebas por parte de los psicólogos.
3. Identificar los principios éticos esenciales que rigen el uso de pruebas por parte de los psicólogos.
4. Describir el sistema de tres niveles de los requisitos que debe cumplir el usuario de pruebas.
5. Identificar las tres fuentes de las "leyes".
6. Resumir las principales estipulaciones de estas leyes pertinentes en el campo de las pruebas: Ley de Derechos Civiles, IDEA, ADA, FERPA, Directrices EEOC y la Decimocuarta Enmienda.
7. Identificar los principios importantes implicados en varios procesos judiciales.
8. Describir las principales aplicaciones forenses de las pruebas.
9. Bosquejar las principales generalizaciones acerca de la conexión del campo de las pruebas con la ley.

Ética y ley

En este capítulo, exploramos los aspectos éticos y legales del campo de las pruebas, que tienen una relación muy estrecha, pero no son idénticos. La ética se ocupa de lo que uno debe o no debe hacer de acuerdo con principios o normas de conducta, mientras que la ley se ocupa de lo que uno está obligado o no a hacer de acuerdo con los dictados legales. En muchas áreas, ética y ley se superponen, pues las leyes a menudo se elaboran a partir de nuestras nociones de los principios éticos: es ilegal y antiético asesinar o robar. Sin embargo, ética y ley no son sinónimas; consideremos algunos ejemplos para ilustrar la diferencia. Puede haber una ley local que prohíba la construcción de un edificio a menos de 15 m de una carretera, pero alguien lo hace a 14.70 m. ¿Ilegal? Sí. Por ese motivo, la persona podría ser multada, encarcelada u obligada a mover el edificio. ¿Antiético? Tal vez no, excepto en la medida en que tenemos la obligación de respetar las leyes en general. De hecho, la persona pudo haber pensado que estaba respetando la distancia de 15 metros, pero se equivocó en la medición de la distancia, en cuyo caso su conducta, sin duda, no puede considerarse antiética, aunque sea ilegal. Sin embargo, sería antiético si esa persona vende el edificio a alguien más sin revelar las violaciones a la ley en que incurrió. Si tú le mientes a tu cónyuge acerca de tus ingresos por los honorarios de tus consultas, ¿eso es antiético? Sí. ¿Ilegal? No. Y si le mientes al servicio de administración tributaria, ¿es antiético? Sí. ¿Ilegal? También. Estos sencillos ejemplos ilustran la diferencia entre los aspectos éticos y legales.

En la siguiente sección, examinamos los principios éticos aplicados al campo de las pruebas: deber y no deber. Después, examinaremos las cuestiones legales: tener la obligación de y tener la obligación de no; lo que se debe y lo que no. Considerar estos aspectos requiere cierto grado de madurez en el campo. Encontraremos que muchos principios éticos y leyes se relacionan con temas como confiabilidad, validez y normas, que tratamos en los primeros capítulos. Ésta es la razón de que hayamos reservado las consideraciones éticas y los temas legales para el capítulo final.

Cuestiones éticas «404a

Para empezar nuestra exploración de la ética aplicable al uso de las pruebas, consideremos estos casos:

- El Dr. Nina, psicólogo escolar, evaluará a Jim, un niño fastidioso y tal vez con muchos problemas de 12 años de edad, que fue enviado por el director de la escuela. El Dr. Nina evaluará la inteligencia de Jim, tal vez con el WISC, y sus características de personalidad mediante diversos inventarios. Desde luego, el Dr. Nina también entrevistará a Jim, hablará con sus maestros y revisará su expediente escolar. ¿El Dr. Nina debe decir a Jim el propósito de la evaluación? ¿O sería mejor mantenerlo oculto? ¿Qué le dirá el Dr. Nina a los padres de Jim antes de la evaluación? ¿Y después? ¿El Dr. Nina debe informar a los padres acerca de todos los resultados? ¿O sería mejor que no supieran todo?
- Psychological Test Resources, Inc. (PTR), una editorial de pruebas psicológicas, tiene un instrumento bien establecido para evaluar los déficits de atención. El manual de la prueba ofrece un excelente conjunto de datos sobre su validez y confiabilidad, así como normas actualizadas. PTR acaba de elaborar una versión de la forma de lápiz y papel de la prueba para aplicarse en computadora. ¿Qué estudios debe realizar PTR antes de lanzar la nueva versión?
- El Dr. Mark emplea el informe interpretativo hecho por computadora del MMPI-2. ¿De quién es la responsabilidad de asegurar que el informe sea interpretado de manera válida: del Dr. Mark o de las personas que prepararon el informe? ¿Quién debe preocuparse por esto?

Estos casos ilustran algunos de los aspectos éticos implicados en el uso de pruebas psicológicas. Observemos el uso de la palabra “deber” en estos casos; ella indica que los principios éticos podrían estar implicados. En las siguientes secciones, exploramos los intentos de los psicólogos por desarrollar principios que puedan aplicarse en estos y otros casos.

Antecedentes de la ética profesional

La ética aplicable al campo de las pruebas, que es nuestro principal interés, forma parte del campo general de la ética del psicólogo. Esta ética no existe en el vacío. Hay tres documentos, desarrollados en los campos de la práctica e investigación médica, que proporcionan los antecedentes de los códigos éticos contemporáneos en psicología. Ellos son el origen de conceptos básicos, incluso de la terminología de los códigos específicos de la psicología que examinaremos.

El primer documento clave es el juramento hipocrático. Hipócrates, a menudo considerado como el padre de la medicina, fue un médico griego que realizó su práctica

alrededor del año 400 a. de C. Su juramento, repetido por los médicos a lo largo de los siglos, incorporó las nociones de la primacía del bienestar del paciente, la competencia (y sus límites) y la confidencialidad. Éste es el origen de la frase citada a menudo "no hagas daño". El segundo documento clave es el *Código Nuremberg* (véase U.S. Government Printing Office, 1949). Shuster (1997) consideró este código como "el documento más importante en la historia de la investigación médica" (p. 1436). El código se elaboró a partir de los juicios de Nuremberg después de la Segunda Guerra Mundial, en los que se enjuició a varios grupos de criminales de guerra nazis. El "juicio de los médicos" tuvo que ver con doctores que utilizaron enfermos de los campos de concentración para hacer experimentos médicos. Aunque el contexto del código Nuremberg fue la experimentación más que la práctica ordinaria, este código ha influido en nuestro pensamiento acerca de la práctica del cuidado de la salud. En donde el código hable de "experimentación", podemos pensar en "práctica" (o terapia o evaluación) y tendremos los puntos esenciales de los códigos éticos actuales. El Código Nuremberg ha tenido una influencia especial por su insistencia en el principio de consentimiento informado, un desarrollo importante que se suma al juramento hipocrático. El tercer documento clave es el *Informe Belmont*, creado en 1979 con financiamiento del Departamento de Salud, Educación y Bienestar de EUA (USHEW, 1979). Este informe establece los principios éticos básicos que rigen el uso de humanos en la investigación. Aunque el contexto del informe fue la investigación más que la práctica, los principios articulados en el informe han influido en los códigos de ética profesionales en la práctica. El informe identifica tres principios éticos básicos: respeto por las personas, beneficencia y justicia. Veremos estos principios en los códigos éticos del psicólogo.

¡Inténtalo!

Revisa la página de internet del Center for Study of Ethics in the Professions: http://ethics.iit.edu/ Contiene códigos éticos de numerosas asociaciones profesionales, de negocios, etc. ¿Puedes encontrar un código ético de organizaciones relacionadas con la psicología? (Pista: Haz clic en Codes of Ethics Collections y busca Mental Health and Counseling.) También puedes encontrar el juramento hipocrático en este sitio.

Los códigos éticos de profesiones específicas tienen distintos propósitos. El primero y más importante es ofrecer una guía. La mayoría de los profesionales de manera consciente acepta principios como "haz el bien y evita el mal" y "no hagas daño". Sin embargo, dentro de cualquier profesión, hay situaciones comunes donde los principios más generales no ofrecen una guía suficiente, por lo que el profesional necesita estar sensibilizado a estas situaciones y recordar cuál es la mejor idea acerca de la conducta apropiada. Como parte de esta función de sensibilización, los códigos éticos a menudo ayudan a lidiar con principios en conflicto, es decir, indican cuál de estos principios tiene preponderancia en una situación particular.

La segunda razón de los códigos éticos profesionales es proteger la reputación de la

profesión. En este caso, el código representa una especie de contrato social entre la profesión y la sociedad. En cuanto a esto, muchas asociaciones profesionales tienen procedimientos para clasificar las quejas por prácticas poco éticas y procedimientos para evaluarlas. La base de una queja es una presunta violación al código ético. En este sentido, el código no es sólo una guía para realizar una buena práctica, sino también un documento cuasi legal aplicable a los miembros de la asociación.

Fuentes de los principios éticos de la evaluación

Hay dos documentos principales de los principios éticos aplicables a la evaluación psicológica. El primero es el *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2013). Por lo general, se denomina simplemente "el *Standards*"; nos referimos a este documento con amplitud en los capítulos 3 al 6. Recordemos que el *Standards* está lleno de referencias a lo que se debe o no hacer en la construcción, aplicación e interpretación de pruebas. Estos "deber" y "no deber" constituyen los principios éticos de la profesión.

El cuadro 16-1 enumera las secciones del *Standards*. Nuestro tratamiento en los capítulos 3 al 6 se enfocó principalmente en los primeros capítulos: cuestiones relacionadas con validez, confiabilidad, normas, construcción de pruebas y neutralidad. Los capítulos 8 y 9 abordan los "derechos y responsabilidades" de las personas que contestan y aplican pruebas, respectivamente. Los últimos capítulos retoman temas especiales que surgen en contextos particulares, por ejemplo, en la evaluación para contratar un empleado o en la escuela.

Cuadro 16-1. Bosquejo del Standards for Educational and Psychological Tests

Introducción
1. Validez
2. Confiabilidad/precisión y errores de medición
3. Neutralidad en las pruebas
4. Diseño y elaboración de pruebas
5. Puntuaciones, escalas, normas, puntuaciones de corte y puntuaciones ligadas
6. Aplicación, calificación, informe e interpretación de las pruebas
7. Documentación de apoyo a las pruebas
8. Derechos y responsabilidades de quienes contestan una prueba
9. Derechos y responsabilidades de usuarios de pruebas
10. Aplicación de pruebas y evaluación psicológica
11. Aplicación de pruebas en contextos laborales y acreditación
12. Aplicación de pruebas y evaluación educativa
13. Usos de las pruebas en evaluación de programas, estudios de políticas públicas y responsabilidad

La segunda fuente importante de principios éticos aplicables a las pruebas es *Ethical Principles of Psychologists and Code of Conduct* [Principios éticos del psicólogo y

código de conducta] (APA, 2002) de la American Psychological Association. Varias fuentes abrevian este título difícil de manejar como "Principios éticos" o "Código de conducta". El documento mismo emplea el término *código ético*, que adoptaremos aquí. El primer código que publicó la APA fue en 1953; éste se ha revisado en varias ocasiones, las más recientes son de 1992 y 2002. El código de 2002 constituye una revisión sustancial del de 1992 tanto en organización como en terminología, pero no en lo esencial. La APA hizo leves modificaciones en la edición de 2010, lo que ha llevado a referencias confusas a veces al código de 2002 y a veces al de 2010. Nosotros nos referiremos al código de 2002.

El cuadro 16-2 presenta un esquema de la versión de 2002 del **Código ético de la APA**. Después de la introducción y el preámbulo, el código presenta cinco principios básicos. En el cuadro 16-3 aparece la primera oración de cada principio para indicar su contenido básico.

Cuadro 16-2. Esquema del Código ético de la APA de la edición 2002

Introducción y aplicabilidad
Preámbulo
Principios generales
- **A.** Beneficencia y no maleficencia
- **B.** Lealtad y responsabilidad
- **C.** Integridad
- **D.** Justicia
- **E.** Respeto a los derechos y la dignidad de las personas

Estándares éticos
- **1.** Resolver cuestiones éticas
- **2.** Competencia
- **3.** Relaciones humanas
- **4.** Privacidad y confidencialidad
- **5.** Publicidad y otras afirmaciones públicas
- **6.** Archivo de expedientes y honorarios
- **7.** Educación y capacitación
- **8.** Investigación y publicación
- **9.** Evaluación
  - **9.01** Bases para la evaluación
  - **9.02** Usos de la evaluación
  - **9.03** Consentimiento informado en la evaluación
  - **9.04** Publicación de los datos de las pruebas
  - **9.05** Elaboración de pruebas
  - **9.06** Interpretación de los resultados de la evaluación
  - **9.07** Evaluación hecha por personas no calificadas
  - **9.08** Pruebas obsoletas y resultados de pruebas anticuadas
  - **9.09** Servicios de calificación e interpretación de pruebas
  - **9.10** Explicación de los resultados de la evaluación
  - **9.11** Mantenimiento de la seguridad de las pruebas
- **10.** Terapia

Fuente: *American Psychological Association* (2002).

Cuadro 16-3. Principios generales del Código ético de la APA

Beneficencia y no maleficencia.[1] "El psicólogo procura beneficiar a las personas con quienes trabaja y tiene cuidado de no causarles daño."
Lealtad y responsabilidad. "El psicólogo establece relaciones de confianza con las personas con quienes trabaja."
Integridad. "El psicólogo busca promover la exactitud, honestidad y veracidad en la ciencia, enseñanza y ejercicio de la psicología."
Justicia. "El psicólogo reconoce que la neutralidad y justicia dan derecho a todas las personas a acceder y beneficiarse de las contribuciones de la psicología."
Respeto a los derechos y la dignidad de las personas. "El psicólogo respeta la dignidad y el valor de todas las personas y sus derechos a la privacidad, confidencialidad y autodeterminación."

Fuente: *American Psychological Association* (2002), *Ethical Principles of Psychologists and Code of Conduct*.

Las secciones numeradas del código contienen los "estándares", cada uno de los cuales consta de varios puntos específicos. Indicamos los títulos de cada punto sólo de la sección 9: Evaluación, por su evidente pertinencia para nuestro tema. Sin embargo, otras secciones también contienen puntos cruciales aplicables a las pruebas. De especial importancia, como veremos, son las secciones sobre competencia, relaciones humanas, privacidad y confidencialidad, y archivo de expedientes.

Otras fuentes

Además del *Standards* y el Código ético de la APA, hay otras fuentes pertinentes para el uso ético de las pruebas. La APA publica de manera periódica otras declaraciones sobre normas de conducta pertinentes para aspectos particulares de la aplicación de pruebas. Dichas declaraciones aparecen a menudo en *American Psychologist*, la revista insigne de la APA, y se pueden consultar de manera práctica en la página de internet del departamento de ética de la APA (www.apa.org/ethics/). Un ejemplo proviene del Joint Committee on Testing Practices [Comité mixto de la práctica en el campo de las pruebas] (JCTP), que publicó un código, *Fair Testing Practices in Education* [Aplicación neutral de pruebas en la educación] (JCTP, 2004), así como otras declaraciones sobre el uso de pruebas. JCTP, sin embargo, se disolvió en 2007.

Otras asociaciones profesionales –como el *American Counseling Association* [Asociación Americana de Orientación] (ACA) y el National Association of School Psychologists [Asociación nacional de psicólogos escolares] (NASP)– tienen sus propios códigos de ética, que incluyen referencias específicas al uso de pruebas. En general, los códigos de la APA, ACA y NASP tienen una estructura muy similar y una gran congruencia entre sí. Sus diferencias tienen que ver primordialmente con el lenguaje y el

contexto. El Buros Center for Testing, mejor conocido como la editorial del *Mental Measurements Yearbooks*, tiene una útil lista de 14 códigos éticos relacionados con el uso de pruebas; véase http://buros.org/standards-codes-guidelines. Otra lista muy extensa de códigos éticos pertinentes para la psicología de diversos países es la excelente compilación de Ken Pope, que se puede consultar en http://kspope.com/ethcodes/index.php

¡Inténtalo!

Consulta el Código ético de la APA en www.apa.org/ethics. ¿Qué dice el código en la sección 9.3 acerca del consentimiento informado?

Por último, señalamos que una gran cantidad de libros brinda explicaciones de los códigos éticos pertinentes en el campo de las pruebas (así como de otras cuestiones profesionales). Entre ellos se encuentra *Ethics in Psychology and the Mental Health Professions: Standards and Cases* (Koocher & Keith-Spiegel, 2008), que ofrece un tratamiento extenso de los temas éticos y ejemplos interesantes. El capítulo 9, Evaluación psicológica: tribulaciones de las pruebas, es de especial relevancia para nuestro tema. *Decoding the Ethics Code: A Practical Guide for Psychologists, Updated* (Fisher, 2010) es otro ejemplo de dichos libros.

Generalizaciones acerca del uso ético de las pruebas

De todas estas fuentes, derivamos los siguientes principios éticos esenciales aplicables al uso de las pruebas. Este resumen (véase Resumen de puntos clave) no sustituye al texto completo de estos códigos, a cuya lectura alentamos al lector para conocer más detalles. Sin embargo, este resumen cubre los principales puntos de un tratamiento introductorio. Nuestro principal punto de referencia es el Código ético de la APA.

Competencia

Para utilizar las pruebas con responsabilidad, el psicólogo debe desarrollar su **competencia** en los conceptos y metodología de la evaluación. Los conceptos incluyen las ideas básicas revisadas en los capítulos 3 al 6: normas, confiabilidad, validez y elaboración de pruebas, incluyendo neutralidad. La metodología incluye los procedimientos específicos aplicables a una prueba específica, como aplicación, calificación y características técnicas. La competencia respecto de la metodología implica seguir los procedimientos de aplicación, calificación, etc. Una persona que carece de los conceptos y la metodología de la evaluación no debe usar pruebas.

El principio de competencia tiene varios principios subsidiarios. El psicólogo es responsable de actualizar de manera continua su conocimiento y habilidades relacionados con la evaluación y debe reconocer los límites de su competencia. (El término límites de la competencia es común en la literatura ética.) Por ejemplo, un psicólogo puede ser competente en el uso de pruebas individuales de inteligencia, pero no en el de las técnicas proyectivas; o puede ser competente en el uso de una técnica proyectiva (digamos, las frases incompletas de Rotter), pero no en otra (digamos, el sistema de Exner del Rorschach). El psicólogo debe reconocer estos límites de su competencia. Por último, el código pone especial atención en la necesidad de conocimiento acerca de diversas poblaciones; el psicólogo necesita estudiar y ser sensible a factores como antecedentes culturales y preferencias de lenguaje.

Consentimiento informado

El paciente, cliente o sujeto debe dar su consentimiento de manera voluntaria para la evaluación, por lo que el psicólogo es responsable de informar a la persona acerca de la naturaleza y propósito de ésta.

Además, el psicólogo debe ofrecer esta información en una forma y lenguaje comprensible para la persona. En el caso de niños u otras personas con capacidades limitadas, se debe obtener el consentimiento de un padre, un tutor legal o un sustituto

apropiado. El **consentimiento informado** supone que la persona puede cancelarlo en cualquier momento.

El Código Ético de la APA incluye algunas excepciones al principio de consentimiento informado. En el caso de una evaluación ordenada legalmente, el psicólogo debe informar a la persona acerca de la naturaleza y propósito de la evaluación a pesar de que no requiera su consentimiento. También hay casos de consentimiento implícito; por ejemplo, una persona que solicita un trabajo da su consentimiento implícito, pues al buscar empleo la persona da a entender su buena disposición para realizar el proceso de solicitud al trabajo, incluyendo cualquier prueba. Los programas de evaluación institucional, por ejemplo, en la escuela, también constituyen una excepción.

Conocimiento de resultados

El paciente, cliente o sujeto tiene derecho a conocer todos los resultados de las pruebas. De ahí que el psicólogo deba proporcionar estos resultados en un leguaje que sea razonablemente comprensible para el individuo. Este principio, como el de consentimiento informado, surge de considerar al cliente como un individuo autónomo y autodeterminado. El principio de **conocimiento de resultados**, así como el de consentimiento informado, tiene ciertas excepciones, por ejemplo, en las pruebas para contratar empleados.

Confidencialidad

El psicólogo debe tratar los resultados de las pruebas como información confidencial. Sólo se pueden dar a conocer a otros profesionales calificados *y* con el consentimiento del cliente. Desde luego, el psicólogo no debe referirse a estos resultados fuera del contexto y el propósito para el que se obtuvieron.

Un principio subsidiario importante de la **confidencialidad** se relaciona con el archivo de expedientes. El psicólogo debe mantener los resultados de la evaluación de una manera segura y, con el tiempo, disponer de ellos de un modo que compagine la confidencialidad de la información con una retención discreta de documentos. Hay excepciones al principio de confidencialidad, como cuando la ley o un juicio requieren la información.

Aquí hay un punto importante que no tiene una relación directa con el campo de las pruebas, pero que vale la pena mencionar. La responsabilidad general del psicólogo de mantener la confidencialidad pasa a segundo término cuando puede ocurrir un grave daño al cliente o a otras personas. Por ejemplo, si el psicólogo se entera, mediante las pruebas u otros medios, de que un individuo planea asesinar a alguien o suicidarse, el principio más general de evitar el daño tiene prioridad sobre el de confidencialidad. Consulta *Tarasoff contra University of California* (P.2d 334, 9th Circuit 1976) para conocer los detalles de un caso de referencia sobre este punto.

Seguridad de las pruebas

El psicólogo debe mantener la **seguridad de las pruebas**. Los materiales de las pruebas se mantienen en un ambiente seguro; los reactivos no se revelan en conversaciones casuales o medios públicos. Se hacen excepciones en los programas de formación y cuando lo ordena la ley.

Los cinco principios que se bosquejaron aquí cubren los puntos principales de los códigos éticos relacionados con las pruebas. Sin embargo, hay tres puntos adicionales con una aplicación más restringida. Ahora los consideraremos.

Elaboración y publicación de pruebas

Los otros principios que bosquejamos aquí se aplican a las pruebas que emplean los psicólogos. Muchos de ellos las usan, pero relativamente pocos elaboran nuevas pruebas. Los psicólogos que se dedican a la construcción y publicación de pruebas tienen obligaciones especiales. Deben mantener altos estándares en la elaboración y abstenerse de hacer afirmaciones injustificadas acerca de la calidad de sus productos. Se espera que los creadores de pruebas sean expertos en temas como confiabilidad, validez y normas, y que apliquen sus conocimientos en el proceso de elaboración de pruebas. Después de que se elabora una prueba, los psicólogos deben proporcionar información adecuada sobre las características técnicas de la prueba; además, deben tener la debida precaución al describirla.

Sistemas de calificación/interpretación automatizada

En años recientes, hemos sido testigos de la proliferación de los sistemas de calificación e interpretación automatizadas, sobre todo tratándose de las pruebas más usadas. Los informes narrativos generados por computadora son un buen ejemplo de ello. Hemos examinado algunos en capítulos anteriores. Los códigos éticos para psicólogos expresan un interés especial en estos sistemas; el punto esencial es que el psicólogo que emplea estos sistemas conserva la responsabilidad de la propia interpretación de los resultados de la prueba. La responsabilidad no se transfiere al creador del sistema, aunque, como en la sección anterior, los creadores tienen sus propias responsabilidades éticas.

Personas no calificadas

El psicólogo no permite o condona el uso de pruebas por parte de personas no calificadas. Por ejemplo, en su ejercicio profesional, el psicólogo no permite que un miembro de su equipo de trabajo con funciones de oficina y sin preparación para usar las

pruebas aplique el PPVT o el WISC a sus clientes. Los psicólogos también toman medidas prudentes cuando observan que una persona no calificada aplica pruebas incluso en circunstancias más allá de su control inmediato.

Requerimientos de los usuarios de pruebas «409a

La profesión de la psicología ha intentado controlar el uso de las pruebas especificando los requerimientos para compararlas. Solemos referirnos a estos intentos bajo el encabezado de **requerimientos de los usuarios de pruebas**. Un título más exacto sería "requerimientos para comprar pruebas". Al menos es posible, en el caso de las pruebas publicadas para su comercialización, ejercer cierto control en el punto de venta. Una vez que se vende, es muy difícil controlar su uso. Uno de los primeros intentos de la APA para crear un código ético se relacionó precisamente con este punto (Eyde *et al*., 1988). Este esfuerzo temprano pone de relieve el papel central que han tenido las pruebas en la psicología y la preocupación especial que los psicólogos han tenido desde hace mucho tiempo respecto del uso apropiado de las pruebas. Como Koocher y Keith-Spiegel (1998) señalaron, "Las pruebas psicológicas están en el corazón mismo de la historia de nuestra profesión" (p. 166).

En el capítulo 1, hicimos un repaso de la historia del *Standards for Educational and Psychological Tests*, un documento citado en repetidas ocasiones en este número. La edición de 1950 del *Standards* introdujo un sistema de tres niveles para los requerimientos de los usuarios de pruebas. Las ediciones posteriores abandonaron este sistema; sin embargo, muchas editoriales de pruebas encontraron en este sistema un mecanismo práctico para el molesto tema de quién puede o no comprar varias clases de pruebas.

Muchas editoriales adoptaron el sistema y siguen usándolo aunque ya no se encuentre en el *Standards*.

Los tres niveles de este sistema son:

Nivel A – Se requiere un nivel mínimo de formación. La aplicación de la prueba implica leer instrucciones sencillas. Incluye pruebas como las de aprovechamiento educativo y eficiencia en el trabajo.

Nivel B – Se requiere cierto conocimiento de las características técnicas de las pruebas. Incluye pruebas como las de capacidad mental de aplicación grupal e inventarios de intereses.

Nivel C – Se requiere una formación avanzada en la teoría de las pruebas y en las áreas correspondientes de contenido. Incluye las pruebas de inteligencia de aplicación individual y las pruebas de personalidad.

En años recientes, muchas editoriales han agregado una categoría dirigida a los profesionales del cuidado de la salud, por ejemplo, terapeutas ocupacionales y enfermeras, para permitir el acceso a un rango restringido de productos de las categorías B y C.

¡Inténtalo!

¿Cómo clasificarías las siguientes pruebas en el sistema de tres niveles para los requerimientos de los usuarios de pruebas?

Stanford Achievement Test _____	WAIS _____
Frases Incompletas de Rotter _____	Strong Interest Inventory _____

Resumen de puntos clave 16-1

Resumen de los principios éticos relacionados con las pruebas «410a
Principios de aplicación amplia:

- Asegurar la competencia
- Obtener el consentimiento informado
- Dar a conocer los resultados
- Mantener la confidencialidad
- Resguardar la seguridad de las pruebas

Principios de aplicación más restringida:

- Establecer estándares altos para la elaboración de pruebas
- Asumir la responsabilidad de los informes automatizados
- Trabajar para evitar el uso de pruebas por parte de personas no calificadas

Aunque este sistema de tres niveles es muy conocido, su aplicación y ejecución reales parecen ser un poco irregulares (Robertson, 1986). Un consorcio de asociaciones profesionales financió un proyecto para desarrollar un método más sofisticado y basado en datos para comprobar que el usuario cumpla con los requerimientos (Eyde *et al.*, 1988; Moreland *et al.*, 1995). El proyecto se llevó a cabo, pero este método no parece haber tenido mucho impacto en la práctica real.

Aspectos legales

¿Qué tienen que ver las pruebas con la ley? ¿De qué modo los temas aparentemente arcanos de los coeficientes de validez y los grupos de estandarización pueden relacionarse con la esencia de las legislaturas y los tribunales? Como veremos, bastante. Consideremos las siguientes preguntas.

- La ABC *Widget Company* emplea el Pennsylvania Screening Test como auxiliar para contratar empleados. Un aspirante afroamericano (que no fue contratado) presenta una denuncia ante la Comisión de Igualdad de Oportunidades de Empleo porque la prueba fue discriminatoria. ¿Qué principios legales se aplican a esta denuncia?
- Un psicólogo escolar emplea el WISC para determinar si un asiático-americano debe ser asignado a un programa de educación especial. ¿Eso es legal?
- En el último caso, los padres del niño quieren conocer los resultados. ¿El psicólogo está obligado legalmente a dar los resultados a los padres?
- Un estudiante asegura que necesita tiempo extra para terminar una prueba debido a dificultades de aprendizaje. ¿La escuela está obligada legalmente a complacer esta petición?
- Eres psicólogo clínico, y un juez te pide determinar si una persona acusada de asesinato es "competente para ser juzgada". ¿Qué harías? ¿Usarías alguna prueba?

Éstos son sólo algunos ejemplos de las maneras fascinantes en que las pruebas se conectan con la ley. En las siguientes secciones, exploraremos estos ejemplos. Pero antes de empezar, una nota de precaución. Cuando presentamos pruebas como WAIS y MMPI, a menudo señalamos que el uso real de estas pruebas requiere una formación avanzada. Nuestra exposición fue sólo una introducción y un panorama general. Una precaución similar se aplica a nuestro tratamiento de los aspectos legales: presentamos un panorama general de las leyes y su aplicación, pero esto no te dará la preparación necesaria para presentar un caso en un juicio.

Nuestra presentación de temas legales, leyes aplicables y procesos judiciales se ubica en el contexto de EUA. Evidentemente se deben hacer cambios cuando se consideren leyes y sistemas legales de otros países.

Áreas de aplicación: panorama general

Las pruebas se conectan con los temas legales en cuatro áreas (véase Resumen de puntos clave 16-2). La primera es el uso de pruebas para decisiones de empleo, que significa contratar, despedir, promover o áreas afines. El principal interés es la igualdad de oportunidades por raza y género. En la práctica, la contratación ha recibido la mayor

atención.

La segunda área es la *educación*, que se divide en dos subcategorías de actividades. La primera es el uso de las pruebas para tomar decisiones de ubicación, sobre todo en los programas de educación especial. La segunda subcategoría se relaciona con la certificación, por ejemplo, para recibir el diploma de bachillerato.

En el caso de las áreas de empleo y educación, ha habido dos grupos meta principales: las minorías raciales/étnicas y las personas con discapacidad. Las minorías raciales/étnicas, como las define el gobierno de EUA, son los afroamericanos, hispanos o latinos, asiático-americanos y americanos nativos.[1] Definiremos la categoría de personas con discapacidades más adelante en relación con los *Americans with Disabilities Act* [Ley de Estadounidenses con Discapacidades] (ADA).

La tercera área es la **psicología forense,** que se ocupa de la aplicación de la psicología, incluyendo la de las pruebas, en acciones legales. Algunos psicólogos se especializan en actividades forenses. Aquí usamos el término en un sentido lato para referirnos a cualquier acción del psicólogo que se lleve a cabo en los procesos de los tribunales. La corte misma puede llamar a un psicólogo para dar testimonio experto en temas como la competencia mental. Además, el psicólogo puede servir como testigo experto de la parte acusadora o de la defensa; por ejemplo, en el caso de una lesión en la cabeza sufrida en un accidente, un neuropsicólogo puede atestiguar acerca del grado del daño con base en los resultados de varias pruebas. Examinamos ejemplos pertinentes en el capítulo 10.

La cuarta área de aplicación es la de los programas autorizados de evaluación, sobre todo los de evaluación estatal. Discutimos estos programas en el capítulo 11, por lo que no es necesario extendernos aquí, excepto para señalar que estas pruebas son un ejemplo de la conexión entre pruebas y ley.

Resumen de puntos clave 16-2

La conexión entre pruebas y ley: cuatro contextos principales

1. Empleo
2. Educación (ubicación, certificación)
3. Forense
4. Programas autorizados de pruebas

Definición de las leyes

En la conversación cotidiana, la palabra "ley" connota una prescripción escrita que se origina en un cuerpo legislativo. Hay una ley, en la mayoría de los estados, que dice que el límite de velocidad en carreteras rurales interestatales es de 100km/h. Hay una ley que dice que, para comprar bebidas alcohólicas, se deben tener 21 años de edad al menos. Sin embargo, este significado común de la palabra "ley" no es suficiente para nuestros propósitos. Para discutir la conexión de las pruebas y "la ley", necesitamos una definición más amplia. Para nuestros propósitos, las leyes se originan en tres fuentes (véase

Resumen de puntos clave 16-3).

El primer tipo de ley es la **ley estatutaria o legislación**. Éste es el significado común de la palabra **ley;** estas leyes tienen su origen en un cuerpo legislativo, como el Congreso de EUA, la legislatura estatal o el cuerpo del gobierno local. Por lo común, la acción de la legislatura debe ser aprobada por el ejecutivo (p. ej., presidente o gobernador) antes de convertirse en ley. Por sencillez, incluimos las constituciones de EUA y de los estados en esta categoría. El segundo tipo de ley es la **ley administrativa o regulaciones**. Una agencia administrativa, por lo general en la rama ejecutiva del gobierno, prepara estas regulaciones. Muy a menudo, proporcionan los detalles para implementar una ley específica. Aunque no pasan por una legislatura, estas regulaciones suelen tener la fuerza de la ley. El tercer tipo de ley es la **jurisprudencia**, es decir, los fallos de los tribunales. Los tribunales interpretan el significado de las leyes cuando las aplican en circunstancias particulares. El fallo de un tribunal, una vez que ha sido dictado, tiene la fuerza de la ley, al menos dentro de la jurisdicción del tribunal. En nuestra definición de fallo de un tribunal se incluyen los decretos y las órdenes de consentimiento; en estas circunstancias, el tribunal no dicta un fallo en favor de una de las partes contendientes, sino que en el proceso las partes llegan a un acuerdo de cómo resolver un asunto.

La fuerza de la ley es casi igual sin importar su origen: estatutaria, administrativa o jurisprudencial. Es decir, todas estas fuentes dan origen a leyes que se deben obedecer.

Leyes relacionadas con las pruebas

Aquí bosquejamos las leyes estatutarias y administrativas de importancia primordial en el campo de las pruebas. En la siguiente sección, examinaremos algunos fallos pertinentes de tribunales (procesos judiciales). Empezamos con los términos básicos y las fuentes de información.

La mayoría de las leyes pertinentes para el campo de las pruebas es de origen federal. Hay algunas leyes estatales e incluso, a veces, locales que también están relacionadas. Cada estado también suele tener leyes y regulaciones específicas diseñadas para implementar leyes federales dentro de la estructura organizacional del estado. Sin embargo, aquí nos concentramos en las leyes federales, ya que son las fuerzas dominantes relacionadas con el campo de las pruebas.

Referirse a las leyes trae consigo una avalancha de números e iniciales. Para ayudar con nuestra revisión, identificamos los siguientes términos y abreviaturas especiales:

U.S.C. representa *United States Code* [Código de EUA] y significa que una ley pasó por el Congreso de EUA. Un número que identifica el área general de la ley precede a las iniciales U.S.C, y un número de sección sigue a las iniciales. Por ejemplo, 29 U.S.C [§]791 designa parte del *Rehabilitation Act* [Ley de Rehabilitación] que describimos más adelante. El símbolo [§] se lee como "sección".

C.F.R. representa *Code of Federal Regulations* [Código de Regulaciones Federales]; se trata de una ley administrativa, mientras que U.S.C es una ley estatutaria. Los números que preceden y siguen a las iniciales C.F.R. identifican la regulación específica.

Por ejemplo, 29 C.F.R. 1607 designa al EEOC *Uniform Guidelines on Employee Selection Procedures* [Directrices Fijas de Procedimientos de Selección de Empleados EEOC] que describimos más adelante.

P.L. representa *Public Law* [Ley Pública]; una P.L. específica, por ejemplo, P.L. 94-142, tiene dos partes. La primera (94) indica el año del congreso, en este caso el Congreso 94. La segunda parte indica la ley numerada dentro del año del congreso, aquí la ley 142 que pasó por el Congreso 94. El número de P.L. no tiene relación con el del U.S.C., y en algunos casos, como el de nuestro ejemplo, se convierte en la forma popular para referirse a la ley. En otros casos, la denominación popular de una ley proviene de su título, por ejemplo, Civil Rights Act [Ley de los Derechos Civiles] o del acrónimo como en el caso de IDEA: Individuals with *Disabilities Education Act* [Ley de Educación para Individuos con Discapacidades]. En realidad, IDEA es la sucesora de P.L. 94-142: una se conoce por su acrónimo y la otra por su número de P.L.

En la figura 16-1, enumeramos las principales leyes relacionadas con el campo de las pruebas en una secuencia más o menos histórica. Sin embargo se debe hacer hincapié en dos puntos acerca de este orden.

Principales áreas de aplicación	Principales leyes
Todas las áreas	Decimocuarta Enmienda a la Constitución de EUA (1868)
Discapacitados en la educación	*Education of the Handicapped* (1970), P.L. 94-142 (1975), IDEA (1990, 1997, 2004, 2011)
Empleo	Ley de Derechos Civiles (1964), Directrices EEOC (1975), Ley de Derechos Civiles Revisada (1991)
Programas educativos	*Elementary and Secondary Education Act* (ESEA, 1965), *No Child Left Behind* (2002)
Discapacidades en la vida civil	Ley de Rehabilitación (1973), Ley de Estadounidenses con Discapacidades (1990)
Privacidad	Ley de Derechos Educativos y Privacidad de la Familia (1974), HIPPA (*Health Insurance...*, 1996)

Figura 16-1. Principales leyes relacionadas con el campo de las pruebas.

Primero, muchas de estas leyes están en continua revisión. Ésta ocurre a menudo por medio de enmiendas en las que el nombre y el número de la ley se mantienen intactos, pero se agregan las enmiendas. En otras ocasiones, la revisión es tan radical que se justifica un nombre y número nuevos, aunque muchas de las ideas básicas de la nueva ley provienen de la versión anterior. Rastrear el linaje de una ley puede ser un gran desafío; en parte, a causa de estos cambios: lo que una vez fue legal ahora puede ser ilegal y viceversa.

Segundo, entre las leyes descritas, hay muchas referencias cruzadas. Las leyes que se enumeran en la figura 16-1 forman un tejido o red; no son entidades aisladas, independientes por completo.

Por ejemplo, una ley puede indicar que incorpora una definición de otra ley, en cuyo

caso, para interpretarla, el lector debe consultar la otra ley. Aquí hay un punto relacionado importante: los nombres cambian sin que cambie el significado. Por ejemplo, las primeras leyes hicieron referencia a los minusválidos, pero en las más recientes emplean el término discapacitado o persona con discapacidad para identificar la misma categoría. Los primeros documentos también se referían a los negros, mientras que ahora prevalece la palabra afroamericano.

Tercero, muchas "leyes" tienen versiones estatutarias y administrativas, por lo que necesitamos consultar ambas fuentes para comprender cabalmente "la ley". A menudo, las versiones administrativas explican los detalles de la versión estatutaria, a veces con giros sorprendentes.

Las siguientes fuentes, desconocidas para la mayoría de los estudiantes de psicología, serán útiles para los lectores interesados en buscar el texto completo de las leyes y los procesos judiciales. Las leyes (U.S.C.) y regulaciones (C.F.R.) de EUA están disponibles en páginas de internet generales. Además, varias agencias de gobierno tienen sus propias páginas de internet en las que aparecen las leyes y regulaciones que las agencias supervisan. A menudo, estas páginas también incluyen materiales complementarios, por ejemplo, versiones simplificadas de las leyes y programas de capacitación relacionados con la implementación de las leyes. En el caso de los procesos judiciales, hay dos bases de datos estándar: LexisNexis y Westlaw. Cada una posee una lista exhaustiva y con referencias cruzadas de los fallos de los tribunales. Éstas son herramientas esenciales para los abogados y son fuentes muy útiles para los psicólogos interesados en la conexión entre el campo de las pruebas y la ley. La mayoría de las bibliotecas académicas tiene una subscripción a por lo menos una de estas bases de datos.

Resumen de puntos clave 16-3

Tres fuentes de las leyes

1. Ley estatutaria; legislación
2. Ley administrativa; regulaciones
3. Jurisprudencia; procesos judiciales

Fuentes útiles de información sobre las leyes

Para consultar el texto completo de las leyes (códigos) de EUA, entra en: http://uscode.house.gov/search/criteria.shtml

Para consultar el texto completo de las regulaciones de EUA, entra en: http://www.gpoaccess.gov/cfr/index.html

Para consultar procesos judiciales, así como leyes y regulaciones, entra a cualquiera de estas bases de datos: LexisNexis o Westlaw

Decimocuarta Enmienda

La primera ley pertinente para el campo de las pruebas es la **Decimocuarta Enmienda** [Fourteenth Amendment] a la Constitución de EUA, en particular la Sección 1. Una parte de esta sección es la cláusula del "debido proceso", otra es la de "protección igualitaria". El cuadro 16-4 presenta la Sección 1 de la Decimocuarta Enmienda con las palabras clave resaltadas en negritas. La **cláusula de protección igualitaria** es la parte más importante para el campo de las pruebas. Esta enmienda fue ratificada en 1868 y su intención principal fue evitar que los estados (principalmente los antiguos estados confederados) aprobaran leyes que restringieran los derechos de los antiguos esclavos. ¿Quién habría de pensar que una enmienda constitucional aprobada en 1868, en el periodo posterior a la Guerra Civil, se relacionaría con el uso de las pruebas psicológicas en el año 2013? Pero así es. Ésta es la noción clave: si una prueba (o cualquier otra cosa) actúa para restringir de manera arbitraria los derechos (incluyendo las oportunidades) de algunos individuos (ciudadanos), entonces la cláusula de protección igualitaria de la Decimocuarta Enmienda se vuelve pertinente.[2]

Cuadro 16-4. Sección 1 de la Decimocuarta Enmienda a la Constitución de EUA
Toda persona nacida o naturalizada en EUA, y sujeta a su jurisdicción, es ciudadana de EUA y del estado en que resida. Ningún estado podrá crear o implementar leyes que limiten los privilegios o inmunidades de los ciudadanos de EUA; tampoco podrá ningún estado privar a una persona de su vida, libertad o propiedad sin un **debido proceso** legal; ni negar a persona alguna dentro de su jurisdicción una **protección legal igualitaria**. [negritas del autor.]

Cuando desarrolla las subsiguientes leyes, el Congreso de EUA se refiere con frecuencia a la Decimocuarta Enmienda y, a veces, también usa otros puntos de referencia. El primero es el poder para regular el comercio interestatal, otorgado al Congreso en el artículo I, sección 8(3) de la Constitución de EUA: "Para regular el comercio con naciones extranjeras, y entre los distintos estados, y con las tribus indias". El segundo es el control del presupuesto relacionado con programas federales. Si un estado u otra agencia no realizan ciertas acciones, el Congreso niega los fondos a ese estado o agencia. Controlar el presupuesto puede ser un mecanismo enormemente poderoso; sin embargo, la Decimocuarta Enmienda ha sido el principal punto de referencia para estas leyes y los procesos judiciales relacionados con ellas.

Ley de Derechos Civiles de 1964 y 1991

Damos un salto de la era posterior a la Guerra Civil a la mitad de la década de 1960, la era de los movimientos por los derechos civiles en EUA. La legislación de mayor influencia en el campo de las pruebas ocurrida en los tiempos modernos fue la **Ley de Derechos Civiles** de 1964; en especial, el Título VII se ocupa de la discriminación para dar empleo. De hecho, un nombre popular de esta ley fue *Ley de Igualdad de*

Oportunidades en el Empleo. La principal preocupación era que los patrones organizaran sus procedimientos para dar empleo, incluyendo las pruebas, de modo que los afroamericanos quedaran excluidos. Desde luego, las leyes resultantes abarcaron en última instancia a otros grupos minoritarios, incluyendo a las mujeres. La Ley de Derechos Civiles de 1964 dio origen a la Comisión para la Igualdad de Oportunidades en el Empleo (EEOC, siglas en inglés). La EEOC, a su vez, creó las Directrices Uniformes, que describimos más adelante, el ejemplo perfecto de una ley administrativa. El Congreso aprobó una revisión importante de la Ley de Derechos Civiles en 1991; la mayor parte de las disposiciones de 1964 quedó intacta, pero hubo un cambio importante pertinente para el campo de las pruebas: el uso de normas para subgrupos.

Ley de Rehabilitación de 1973 y Ley de Estadounidenses con Discapacidades de 1990

Dos leyes con una estrecha relación son la Ley de Rehabilitación de 1973 (29 U.S.C. [§]701; P.L. 93-112) y la Ley de Estadounidenses con Discapacidades de 1990 (ADA, siglas en inglés; U.S.C. [§]12101; P.L. 101-336). De varias maneras, **ADA** es una revisión y expansión de la Ley de Rehabilitación, aunque esta última sigue vigente. El propósito primario de estas leyes fue ofrecer acceso a los minusválidos, enfocándose al principio en aspectos como las barreras arquitectónicas. ¿Una persona en silla de ruedas puede hacer uso de la biblioteca pública local? Así, el símbolo internacional de "accesible para los minusválidos" es una silla de ruedas.

Sin embargo, estas leyes interpretan el término minusválido o discapacitado en un sentido amplio que incluye no sólo las dificultades físicas sino también las mentales o psicológicas. La definición incluye "dificultades específicas de aprendizaje", lo cual ha llevado a un crecimiento explosivo en el número de personas y de tipos de "barreras" contempladas por la ley. Estas dos leyes, más que cualquier otra, son las que han dado origen al tema de las adaptaciones de las pruebas educativas y psicológicas. Discutimos este tema de manera extensa en el capítulo 6 bajo el tema de la neutralidad de las pruebas (véase pp. 160-169a»).

El cuadro 16-5 presenta la declaración del propósito de ADA. Observa la referencia a la Decimocuarta Enmienda. Las definiciones básicas incluidas en ADA son, en esencia, las mismas que las de IDEA, que aparecen en el cuadro 16-6.

Cuadro 16-5. Propósitos de la Ley de Estadounidenses con Discapacidades
1. Emitir un mandato nacional claro e integral para eliminar la discriminación contra individuos con discapacidades.
2. Proporcionar estándares claros, sólidos, consistentes y ejecutables relacionados con la discriminación contra individuos con discapacidades.
3. Asegurar que el gobierno federal tenga un papel central en la ejecución de los estándares.
4. Invocar el movimiento de la autoridad del Congreso, incluyendo el poder para ejecutar la Decimocuarta Enmienda y regular el comercio, para abordar las principales áreas de discriminación encarada día a

día por las personas con discapacidades.

Cuadro 16-6. Extractos de las definiciones de discapacidad de IDEA

El término niño con discapacidad se refiere a un niño con retraso mental, deterioros auditivos (incluyendo sordera), deterioros en el habla o lenguaje, deterioros visuales (incluyendo ceguera), perturbación emocional seria (denominado en este documento como "perturbación emocional"), deterioros ortopédicos, autismo, lesión cerebral traumática, otros deterioros en la salud o dificultades específicas de aprendizaje; y quien, por una de estas razones, necesita educación especial y servicios relacionados.

El término dificultad específica de aprendizaje se refiere a un trastorno en uno o más de los procesos psicológicos básicos implicados en la comprensión o el uso de lenguaje, hablado o escrito, que se puede manifestar en la capacidad imperfecta para escuchar, pensar, hablar, leer, escribir, deletrear o hacer cálculos matemáticos.

Trastornos incluidos: este término incluye padecimientos como discapacidades perceptuales, daño cerebral, disfunción cerebral mínima, dislexia y afasia de desarrollo.

Trastornos no incluidos: este término no incluye un problema de aprendizaje que resulta primordialmente de discapacidades visuales, auditivas o motrices, del retraso mental, de una perturbación emocional o de desventajas ambientales, culturales o económicas.

La Ley de Tecnologías de Apoyo de 2004 (P.L. 108-364, 29 USC 3001) puede considerarse un adjunto de ADA, aunque también tiene implicaciones importantes para las leyes educativas que presentamos en la siguiente sección. Esta ley busca promover la disponibilidad y uso de tecnologías de apoyo para personas con discapacidades.

Minusválidos/discapacitados en la educación: P.L. 94-142 e IDEA

Una serie de leyes que aparecieron en 1970 se ocuparon de los discapacitados en escenarios educativos. Aquí enumeramos los puntos de referencia y omitimos las diversas revisiones y enmiendas intermedias. La primera fue la Ley para la Educación de los Minusválidos (EHA, siglas en inglés) en 1970. Una revisión importante se tituló Educación para Todos los Niños Minusválidos, aprobada en 1975, cuya etiqueta, P-L-94-142, tuvo mucha aceptación y se sigue usando en la actualidad. Esta ley fue enmendada al menos en cuatro ocasiones (en cada una con un nuevo número P.L.). Hubo otra revisión importante en 1997, titulada Ley para la Educación de Individuos con Discapacidades, y otra en 2004 (IDEA; 20 U.S.C. [§]1400, P.L. 108-446)[3], y se hicieron extensiones a los bebés y niños de 1 año en 2011. El cuadro 16-6 presenta las definiciones de "niño con discapacidad" y "dificultad específica de aprendizaje" de IDEA.

El punto esencial de todas estas leyes fue ofrecer una "educación pública libre apropiada" a todos los niños, con especial atención a los minusválidos/discapacitados. La frase "educación pública libre apropiada" se usó con tanta frecuencia en estas leyes y documentos asociados que recibió su propio acrónimo: FAPE (siglas en inglés). La

palabra clave de esta frase es "apropiada".

¿Qué será apropiado? ¿Es apropiado ubicar al discapacitado en salones de clase separados? ¿Cuáles son los medios apropiados para evaluar las necesidades del estudiante discapacitado? Esta última pregunta da la entrada a las pruebas en este cuadro. Podemos notar en las definiciones del cuadro 16-6 la referencia a padecimientos específicos y sus probables causas. Las pruebas, casi siempre, están implicadas en estas funciones diagnósticas.

IDEA, así como sus predecesores, exige el desarrollo de un "programa individualizado de educación", conocido popularmente como **IEP** (siglas en inglés). Cada niño con discapacidad debe tener un IEP; la ley especifica quién puede participar en la preparación y ejecución del IEP (incluyendo a los padres), los procedimientos para actualizarlo y el papel de las pruebas en el IEP. El cuadro 16-7 muestra algunas de las disposiciones de IDEA ([§]614) en relación con las pruebas.

No hace falta decir que el psicólogo escolar debe conocer detalladamente esta ley. Los psicólogos clínicos en la práctica privada o en agencias extra escolares también son consultados para aplicar pruebas relacionadas con a) la detección de discapacidades y b) la información pertinente para un IEP.

Cuadro 16-7. Ejemplos de afirmaciones de IDEA relacionadas con las pruebas

Al realizar una evaluación, la agencia educativa local deberá:
- usar diversas herramientas y estrategias de evaluación para obtener información pertinente de las áreas funcional, de desarrollo y académica, incluyendo la que proporcionen los padres;
- no usar una medida o evaluación como criterio único; y
- usar instrumentos técnicamente sólidos que puedan evaluar la contribución relativa de los factores cognitivos y conductuales además de los físicos y de desarrollo.

Cada agencia educativa local asegurará que [las evaluaciones y otros materiales de evaluación]:
- sean elegidos y aplicados de una manera que no sea discriminatoria en términos raciales o culturales;
- sean suministrados y aplicados en el lenguaje y la forma que, con mayor probabilidad, produzca información precisa;
- sean usados para los propósitos para los que las evaluaciones o medidas son válidas y confiables;
- sean aplicados por personal con la preparación y el conocimiento necesarios; y
- sean aplicados de acuerdo con cualquier instrucción dada por el productor de dichas evaluaciones.

El examen de los extractos del cuadro 16-7 revela dos puntos importantes. Primero, las afirmaciones son consistentes con los estándares profesionales que hemos subrayado de principio a fin acerca de las cualidades técnicas de las pruebas. Segundo, las afirmaciones son muy específicas en relación con las obligaciones de las personas que eligen y aplican las pruebas. De ahí que una comprensión general de los estándares profesionales que deben cumplir las pruebas no sea suficiente para cumplir con todos los requerimientos legales. Debemos leer la ley y las declaraciones reguladoras pertinentes.

Otras disposiciones de estas leyes exigen la educación de los discapacitados en el "ambiente menos restrictivo". En la práctica, esto significa "integración", es decir, la inclusión de los alumnos con discapacidades en salones de clase regulares en vez de

salones separados de educación especial. Al igual que con muchos aspectos de la ley, hay algunas excepciones a este principio.

FERPA e HIPPA

La Ley de Derechos Educativos y Privacidad de la Familia de 1974 (**FERPA**, siglas en inglés), también conocida como la enmienda Buckley, tiene como propósito principal garantizar que los individuos tengan libre acceso a información sobre sí mismos, que puedan cuestionar la validez de la información de los archivos de las agencias y que otras partes no tengan acceso injustificado a información personal. La pertinencia especial de FERPA para el campo de las pruebas tiene que ver con las puntuaciones: la información de las pruebas tiene que estar disponible para el examinado o, en caso de los menores, para los padres o tutores. Aunque puede ser difícil creerlo en la actualidad, antes de FERPA era común que las escuelas y otras agencias prohibieran a los padres ver los resultados de las pruebas de sus hijos. Además, la ley dispone restricciones específicas acerca de dar tal información a otras partes sin el consentimiento del individuo.

Otra ley federal, aunque con un objetivo específico distinto, tiene casi el mismo efecto que FERPA. La Ley de Responsabilidad y Portabilidad de los Seguros de Salud de 1996 (P.L. 104-91), popularmente conocida como **HIPPA** (siglas en inglés), está dirigida principalmente a la industria de la salud, que incluye la salud mental y el uso de pruebas psicológicas. En general, FERPA e HIPPA son bastante consistentes con los principios éticos relacionados con la confidencialidad que examinamos antes en este capítulo.

Directrices EEOC

Como resultado directo de la Ley de Derechos Civiles de 1964, la Comisión para la Igualdad de Oportunidades en Empleo (EEOC, siglas en inglés, 1978) publicó los *Uniform Guidelines on Employee Selection Procedures* [Directrices Uniformes sobre Procedimientos de Selección de Empleados]. El código de EEOC es 29 C.F.R. [§]1607. Hubo un financiamiento conjunto entre otras agencias federales, cada una usando su propio código. Nosotros las llamaremos **Directrices EEOC** o simplemente las Directrices; son perfecto ejemplo de regulaciones o leyes administrativas, por lo que el término **directrices** es un eufemismo. El documento impone requerimientos muy estrictos a los patrones.

¡Inténtalo!

Para ver el texto completo de las Directrices EEOC, entra en www.eeoc.gov y encuentra 29 C.F.R. [§]1607. Este sitio también contiene muchas de las leyes citadas en esta sección.

Las Directrices EEOC "tenían la intención de establecer una posición federal fija en el

área de prohibir la discriminación en el empleo debida a raza, religión, sexo o país de origen" (29 C.F.R. [§]1607.18). De manera más específica, las Directrices "se aplican a las pruebas y a otros procedimientos de selección que se usan como base para tomar decisiones de empleo" (29 C.F.R. [§]1607.2B).

Las Directrices usan de manera consistente la frase "pruebas y otros procedimientos de selección". Ya que nuestro tema central son las pruebas, en los siguientes párrafos nos referiremos sólo a ellas, aunque todas las afirmaciones se aplican a cualquier procedimiento, dispositivo o requerimiento de selección. Por ejemplo, las Directrices abarcan los requerimientos de un cierto nivel de educación, digamos, certificado de bachillerato o los resultados de una entrevista. También omitimos ciertas excepciones, como la aplicabilidad en áreas alrededor de las reservas indias. Como sucede con otras leyes que tratamos aquí, queremos transmitir la idea central de las leyes, no todos los detalles.

Para alguien que conoce los principios, que revisamos en este libro, de las pruebas psicológicas, las Directrices EEOC son, en gran medida, ordinarias. En su mayor parte se leen como un libro de texto de pruebas psicológicas. El cuadro 16-8 presenta algunos extractos para ilustrar este punto. De hecho, las Directrices afirman que "buscan ser consistentes con los estándares profesionales generalmente aceptados... como los que se describen en *Standards for Educational and Psychological Tests*" (29 C.F.R. [§]1607.5C). Es comprensible que las secciones más extensas de las Directrices se relacionen con la validez. Aunque las frases de las Directrices son conocidas, llama la atención ver la jerga psicométrica en forma de jerga reguladora. Por ejemplo, la referencia al nivel de significancia de .05 de un coeficiente de validez de criterio se convierte en 29 C.F.R. [§]1607.14B5.

Cuadro 16-8. Extractos de las Directrices EEOC

La evidencia de la validez de una prueba y otros procedimientos de selección mediante estudios de validez de criterio debe constar de datos empíricos que demuestren que el procedimiento de selección es predictivo o tiene una correlación significativa con elementos importantes del desempeño en el puesto. (1607.5B)
Los procedimientos de selección deben aplicarse y calificarse en condiciones estandarizadas (1607.5E)
Por lo general, un procedimiento de selección se considera que está relacionado con el criterio, para propósitos de estas directrices, cuando la relación entre desempeño en el procedimiento y desempeño en la medida criterio sea estadísticamente significativa al nivel de .05 (1607.14B5)

He aquí un resumen sencillo de la idea principal de las Directrices: *Si el uso de una prueba para decidir a quién dar un empleo provoca un impacto adverso (definido más adelante), debe haber información de la validez de la prueba para justificar su uso*. Este resumen contiene tres puntos importantes. Primero, en ausencia de un impacto adverso, las directrices no exigen información de validez (aunque el sentido común sin duda lo hace). Segundo, debemos determinar qué define a la información con validez adecuada. En esencia, esto se refiere a una aplicación apropiada y explícita de la validez

de contenido, de criterio y/o de constructo. En otras palabras, debemos demostrar que la prueba es válida para elegir (o promover) a los individuos en el puesto en cuestión. Varias leyes se refieren a esta pertinencia para el puesto como "necesidad de negocios".

Es decir, la operación exitosa del negocio requiere de ciertos requisitos para el puesto. Tercero, debemos definir el impacto adverso.

Impacto adverso y la regla de cuatro quintos

El concepto de impacto adverso es una parte de las Directrices que no surge directamente de la literatura psicométrica, por lo que sus definiciones generales y operacionales merecen atención especial. El **impacto adverso** se refiere a que la prueba (o cualquier otro procedimiento de selección) da por resultado tasas de selección desfavorables por sexo, raza o grupo étnico. La **regla de cuatro quintos** define de qué magnitud debe ser una diferencia en las tasas de selección para que se considere problemática. La regla dice: "Una tasa de selección de cualquier raza, sexo o grupo étnico menor de cuatro quintos (4/5 u 80%) de la tasa más alta será considerada por las agencias federales de ejecución como evidencia de impacto adverso" (29 C.F.R. [§]1607.4). En la práctica, uno debe determinar la tasa de selección de cada grupo y, luego, la tasa entre las tasas. El cuadro 16-9 muestra un ejemplo en el que se manifiesta un impacto adverso para los solicitantes hispanos.

Cuadro 16-9. Ilustración del procedimiento para determinar el impacto adverso

	Grupo			
	Blancos	**Negros**	**Asiáticos**	**Hispanos**
Solicitantes	50	20	10	20
Contratados	29	11	6	6
Tasa de selección (TS)	.58	.55	.60	.30
TS/tasa más alta	.58/.60 = .97	.55/.60 = .92	.60/.60 = 1.00	.30/.60 = .50

No está de más hacer hincapié en que el hallazgo del impacto adverso no significa que el uso de la prueba es ilegal, es decir, que está prohibido por las Directrices, sino que el usuario debe contar con información adecuada sobre la validez para justificar su uso. Por ejemplo, si el patrón ha mostrado que la prueba es un predictor significativo del desempeño en el puesto, el uso de la prueba está permitido. Además, para repetir una afirmación anterior, en ausencia de impacto adverso (digamos, si 10 hispanos hubieran sido contratados), no habría necesidad de demostrar la validez, aunque podríamos preguntarnos por qué el patrón se molestó en usar la prueba.

Obtención de normas de subgrupos: aparición y prohibición

La regla de cuatro quintos causó notable consternación entre patrones y especialistas en recursos humanos. La demostración o incluso la mención del impacto adverso podrían acarrear que se involucrara en la cuestión a un grupo de funcionarios federales. La manera más obvia de tratar con este tema es realizar los estudios de validez necesarios. Sin embargo, otra manera es forzar más o menos tasas equivalentes de selección por cada grupo. Esto se puede lograr tomando como punto de referencia al propio subgrupo en vez de hacerlo con el total de solicitantes o con un grupo de estandarización unitario. Supongamos que los subgrupos difieren de manera considerable en el rasgo medido por la prueba[4] y seleccionamos, digamos, al mejor 20% de solicitantes. El grupo más alto en el rasgo estará sobrerrepresentado entre los seleccionados, pues tal vez la mayoría de ellosprovenga del grupo más alto. A la inversa, el grupo más bajo en el rasgo estará sub-representado. Sin embargo, si seleccionamos el 20% más alto de cada subgrupo, el impacto adverso desaparecerá automáticamente. Éste es un ejemplo de **obtención de normas por subgrupos**; el procedimiento es equivalente a desarrollar una norma local para cada subgrupo. De hecho, se trata de una práctica que se usó en respuesta a las Directrices EEOC.

La Ley de Derechos Civiles de 1991 prohibió específicamente el uso de la obtención de normas por subgrupos, pues fue vista como una forma de discriminación inversa. La ley dice:

> Prohibición del uso discriminatorio de las puntuaciones de pruebas. Será una práctica ilegal de empleo por parte del demandado, en relación con la selección o derivación de solicitantes o candidatos a un empleo o promoción, ajustar las puntuaciones, usar diferentes puntuaciones de corte o alterar los resultados de pruebas relacionadas con el empleo con base en la raza, color, religión, sexo u origen nacional. (42 U.S.C. [§]2000e-2(1))

Podemos notar que las Directrices EEOC no alentaron, ni siquiera mencionaron, la obtención de normas de subgrupos. Sin embargo, la ley revisada prohibió una práctica que intentaba ocuparse de las disposiciones de las Directrices.

No Child Left Behind Act

Aunque su nombre oficial es "*No Child Left Behind Act of 2001*" [Ley Que Ningún Niño se Quede Atrás] (NCLB; P.L. 107-110), esta ley federal se adoptó en 2002. Se trata de un gran éxito: casi 2000 páginas en algunas versiones. Con un programa planeado para 12 años con fechas tope intermedias para su implementación completa, NCLB tiene implicaciones importantes para el mundo de las pruebas. Recordemos del capítulo 11 nuestra discusión sobre la responsabilidad educativa y la educación basada en estándares. NCLB tradujo muchas de estas nociones generales a disposiciones legales muy detalladas. Exigió la especificación de los estándares educativos y una evaluación amplia para determinar que esos estándares se han cumplido. Como lo sugiere su nombre, la ley hizo hincapié en que todos los estudiantes deben demostrar un desempeño adecuado en

la prueba. Además, la ley incluyó requerimientos precisos para que las escuelas informen las puntuaciones al público por distintas categorías de estudiantes, y exigió la demostración de mejoras en las puntuaciones promedio de un año a otro.

En el capítulo 11 nos referimos a algunos efectos de NCLB en nuestra discusión de los programas estatales de evaluación. Aquí agregamos que esta ley ha sido, tal vez, la fuerza impulsora más importante en la educación pública de EUA en los niveles primario y secundario durante la primera parte del siglo XXI. Probablemente, la ley no ha aumentado la cantidad total de pruebas aplicadas, ya que las nuevas pruebas estatales a menudo sólo reemplazan las baterías de aprovechamiento estandarizadas y previamente usadas. Las numerosas y detalladas disposiciones de la ley consumen enormes cantidades de energía y recursos. La comunidad educativa pública ha hecho objeciones considerables respecto de su factibilidad.

La meta de NCLB es que todos los estudiantes (sin excepción de "ningún niño") tengan un nivel competente en 2014, pero parece poco probable que esta meta se alcance en el tiempo establecido. Aunque la NCLB ha sido muy criticada, el Congreso (su creador) se ha negado con firmeza a revisar la ley. Mientras tanto, el ramo ejecutivo (el presidente) ha empezado a emitir "exenciones" para que los estados no tengan que implementar la ley. Seguiremos pendientes de este tema.

Procesos judiciales ilustrativos

Recordemos de nuestra definición de "leyes" que los fallos de los tribunales constituyen el tercer tipo de ley: jurisprudencia. Los tribunales se ocupan de la aplicación de las leyes estatutarias y administrativas, así como de los fallos previos de los tribunales a casos particulares. Hay un acusador que afirma que alguien (una persona, corporación, escuela, etc.) no se apegó adecuadamente a la ley, mientras que una parte acusada (la persona, corporación, etc.) sostiene que se apegó a la ley. Un abogado, por lo general, representa a cada parte, la acusadora y la acusada. El tribunal escucha el caso y falla en favor de una parte. Para el tipo de casos que nos interesan aquí, los jueces, no el jurado, emiten el fallo. Cuando se presentan por primera vez, los casos se suelen identificar por los nombres del acusador y el acusado, en ese orden, por ejemplo, *Hogan contra Ciudad de Scranton*. Aquí, alguien llamado Hogan, el acusador, afirma que la ciudad de Scranton, el acusado, violó la ley. A menudo, hay múltiples acusadores y acusados en un mismo caso, pero la costumbre es reducir el título a nombres únicos.

La mayoría de las leyes relacionadas con las pruebas es de origen federal, por lo que los tribunales federales se encargan de los casos. El magistrado local se ocupa de tu boleto de estacionamiento porque se relaciona con un reglamento local, pero no se encarga de un caso en el que interviene una ley federal.

¡Inténtalo!

¿Tienes interés en los tribunales federales de EUA? Revisa esta página de internet: www.uscourts.gov

Literalmente miles de casos que llegan a los tribunales implican pruebas psicológicas y las leyes antes descritas. Aquí presentamos algunos casos que ilustran la manera en que los tribunales se ocupan de la conexión entre pruebas y la ley. De ese modo, resumimos los puntos de mayor pertinencia inmediata para el campo de las pruebas omitiendo una gran cantidad de fascinantes detalles, por ejemplo, si el caso fue presentado de manera oportuna, si se falló primero de un modo y después se modificó el fallo tras una apelación, etc., es decir, los detalles que a menudo mantuvieron vivo el caso en los tribunales durante decenas de años.

Griggs contra Duke Power

Una compañía exige el certificado de bachillerato y una puntuación aprobatoria en una prueba de inteligencia para la contratación y/o promoción de sus empleados. Sin embargo, los requerimientos ponen en desventaja a los afroamericanos: ¿constituyen, por lo tanto, una violación a la Ley de Derechos Civiles de 1964? Ésta fue la pregunta en el caso *Griggs contra Duke Power* (1971),[5] uno de los primeros y más celebrados casos

después de la aprobación de la Ley de Derechos Civiles de 1964.

La planta de generación de vapor de Duke Power en Draper, Carolina del Norte, cumplía con los requerimientos. Griggs fue el acusador líder de 13 afroamericanos (a los que se denominaba negros en aquel entonces).[6] Los acusadores tenían menor educación y menores puntuaciones en las pruebas en comparación con los blancos, por lo que afirmaron que los requerimientos –tanto el certificado como las pruebas– eran discriminatorios. La compañía sostenía que no había una intención discriminatoria y que los requerimientos ayudaban a mejorar la calidad general de su fuerza de trabajo. Se usaron tres pruebas: una versión del Otis Beta, el *Wonderlic Personnel Test* (un descendiente del Otis Intelligence Scale) y el Bennett Mechanical Comprehension Test. Conforme avanzó el caso, las pruebas parecían haber evolucionado en una "prueba estandarizada de inteligencia general", frase usada en el fallo final. De acuerdo con la Suprema Corte, el empleador no pudo demostrar una relación razonable entre los requerimientos (el certificado y las pruebas) y el desempeño en el trabajo. La Suprema Corte se basó, en parte, en documentos del *Registro del Congreso* respecto de los debates en el Congreso sobre la intención de que hubiera referencias a pruebas elaboradas de manera profesional. En efecto, la Corte concluyó que la intención del Congreso fue que las pruebas no sólo se elaboraran de manera profesional, sino que tuvieran validez para un propósito específico. Desde luego, este punto, tortuosamente desarrollado a principios de la década de 1970, era el corazón de las Directrices EEOC publicadas en 1978. Las pruebas, o incluso el certificado de bachillerato, pueden usarse para tomar decisiones de contratación, pero sólo si se han validado. En este contexto, **validado** se refiere a mostrar una relación entre el dispositivo de selección y el desempeño en el trabajo. Éste es el criterio de "necesidad del negocio".

Debra P. contra Turlington y GI Forum contra TEA

Un estado exige que los estudiantes aprueben una prueba para graduarse del bachillerato. Las tasas de aprobación en la prueba son considerablemente distintas para las minorías y los estudiantes blancos. ¿El requisito de la prueba viola alguna ley federal? De manera más específica, ¿este uso de la prueba viola la Decimocuarta Enmienda, la Ley de Derechos Civiles de 1964 o cualquier otra ley relacionada? Dos casos semejantes, *Debra P. contra Turlington* (1984) y *GI Forum contra TEA* (2000), ilustran el modo en que los tribunales han afrontado esta pregunta.

El **caso Debra P.** ocurrió en Florida en 1979. Debra P. era una estudiante que actuó como acusadora en una acción de clase presentada a nombre de todos los estudiantes negros de Florida.[7] Ralph Turlington era el comisionado de educación del estado, responsable de implementar el programa de evaluación estatal. En 1976, la legislatura de Florida, mediante el Departamento de Educación, empezó a exigir a los estudiantes que aprobaran una prueba funcional de la capacidad de leer y escribir, el *State Student Assessment Test*, Part 2 (SSAT-II), para poder recibir su certificado de bachillerato. La

prueba se aplicó en el otoño de 1977, 1978 y 1979. Los estudiantes negros reprobaron la prueba casi 10 veces más que los blancos. El pleito se presentó en 1979.

Éstos son los puntos importantes acerca del caso *Debra P.* Primero, se permitió al estado usar la prueba como requisito para el certificado. Segundo, el estado tenía que posponer la implementación de este requisito hasta 1982-1983 con el fin de dar apropiadamente el aviso (esto es parte del debido proceso) acerca del requisito. Tercero, una parte crucial de este caso se dirigió a la adecuada demostración de la "validez instruccional" de la prueba; véase en el capítulo 5 la descripción de este concepto. El estado encargó estudios empíricos amplios para demostrar la validez instruccional. La disposición de trabajo correctivo para quienes reprobaran la prueba también fue una consideración importante. Cuarto, en la opinión del tribunal, el solo impacto desproporcionado de la prueba en un grupo no hacía que la prueba fuera ilegal. Quinto, no hubo una demostración clara de una relación causal entre los vestigios de una enseñanza segregada previa y las tasas actuales de aprobación en la prueba. Los programas correctivos fueron importantes en este punto. Además, desde la perspectiva del tribunal, la imposición de la prueba como requisito, en realidad, pudo haber ayudado a corregir dichos vestigios. Ésta es una aplicación interesante del concepto de validez consecuencial, aunque ni el tribunal ni el acusado usaron este término.

En un caso muy similar al de *Debra P.*, un grupo de estudiantes minoritarios, con el GI Forum como principal acusador, llevó el pleito contra la Agencia de Educación de Texas (TEA, siglas en inglés) a la Corte Distrital federal en San Antonio. A principios de la década de 1980, Texas exigió que los estudiantes aprobaran la prueba *Texas Assessment of Academic Skills* (TAAS), así como que cumplieran con otros criterios, para poder graduarse del bachillerato. Los estudiantes afroamericanos e hispanos reprobaron las pruebas (lectura, matemáticas y escritura) con tasas considerablemente más altas que los blancos, de modo que se presentó un impacto adverso.

En enero de 2000, el juez dictaminó que la prueba TAAS era permisible a pesar de las disparidades en las tasas de aprobación. Hubo varios puntos importantes en la opinión del tribunal. Primero, el juez usó la regla de cuatro quintos EEOC, diseñada originalmente para las pruebas de empleo.[8] Transformó "necesidad del negocio" (véase p. 594) en "necesidad educativa". Segundo, se da mucha importancia a la calidad técnica de las pruebas: elaboración, en especial las medidas para detectar sesgos en los reactivos; confiabilidad; y validez curricular, incluyendo "la oportunidad de aprender". En este caso, validez curricular es lo mismo que validez instruccional en el caso *Debra P.* Tercero, y quizá lo más importante, el juez hizo referencia al impacto de la prueba; en particular, señaló que la prueba ayudó a a) identificar y corregir las debilidades y b) motivar a estudiantes, maestros y escuelas. En Phillips (2000) se puede encontrar un tratamiento detallado del GI Forum y sus lecciones.

Larry P. contra Riles, PASE contra Hannon y Crawford contra Honig

Consideramos estos tres casos porque, en conjunto, ilustran la complejidad de los

siguientes fallos de los tribunales. Los casos nos llevan a un fascinante viaje a través de los tribunales. Al igual que *Griggs* y *Debra P.*, en estos casos estuvieron involucrados múltiples acusadores y acusados, apelaciones, arrestos preventivos y otros procesos legales. Aquí sólo presentamos las principales conclusiones.

En **Larry P.** *contra Riles* (1984), Larry P. era un niño afroamericano asignado a una clase para mentalmente retrasados educables (EMR, siglas en inglés) con base principalmente en una prueba de inteligencia de aplicación individual. Los acusadores afirmaron que esto era una violación a la Ley de Derechos Civiles de 1964, la Ley de Rehabilitación de 1973, la Ley de Educación para todos los Niños Minusválidos de 1975 y las disposiciones de protección igualitaria de las constituciones federal y estatal.

La corte concluyó que las pruebas eran discriminatorias y prohibieron su uso en el futuro. Este fallo, quizá más que cualquier otro, dio origen a titulares como "Pruebas de CI encontradas ilegales" y "Pruebas de inteligencia inconstitucionales". Sin embargo, veamos los siguientes puntos de la decisión completa y el razonamiento del tribunal. Primero, el fallo se aplicó sólo a los estudiantes afroamericanos, no a otras minorías de estudiantes (ni a los blancos). Segundo, el fallo sólo se aplicó a las tareas en las clases EMR o a sus equivalentes sustanciales y no prohibió el uso de las pruebas para otros propósitos, por ejemplo, selección para clases de superdotación intelectual. Tercero, un elemento importante en el razonamiento del tribunal fue la conclusión de que las clases EMR no tenían porvenir. En varios sentidos, el fallo fue una acusación más del programa de clases EMR que de las pruebas. Por último, el tribunal concluyó que las pruebas de CI tenían más peso de lo que estaba justificado en vista de los estatutos estatales.

Mientras tanto, otra vez en Chicago, surgió un caso muy similar: *PASE contra Hannon*. Parents in Action on Special Education [Padres en acción sobre la educación especial] (PASE, 1980) presentaron una demanda contra Joseph Hannon, superintendente de las Escuelas Públicas de Chicago, a nombre de los estudiantes afroamericanos asignados a clases de mentalmente minusválidos educables (EMH, siglas en inglés) con base, en parte, en pruebas individuales de inteligencia. En este caso, el juez dictaminó que las pruebas no tenían sesgos culturales, a excepción de algunos reactivos, los cuales no influían de manera significativa en la ubicación final del estudiante, por lo que las pruebas eran permisibles. El juez también señaló que los psicólogos implicados en la ubicación de alumnos parecían tener la formación profesional adecuada y emplearon un buen criterio profesional. Información adicional además de la de las pruebas de inteligencia también tuvo un papel importante en la ubicación.

En ***Crawford*** *contra Honig* (1994), Demond Crawford era una estudiante afroamericana diagnosticada con dificultades de aprendizaje. Sus padres quisieron que se le aplicara una prueba de CI, pero el estado presentó una objeción citando el fallo del caso *Larry P*. Es interesante que el mismo juez que llevó el caso *Larry P*. dictaminó en sentido contrario de (renunció a) su primer dictamen, de modo que permitió el uso de una prueba de CI con Crawford y, de hecho, con todos los estudiantes afroamericanos.

La lectura cuidadosa de los procesos de estos casos revela cuatro puntos importantes que no fueron tratados en los dictámenes finales, pero que fueron parcialmente

determinantes. Primero, la diferencia entre el diagnóstico del retraso mental y las dificultades de aprendizaje fue crucial, y las pruebas pueden ser de especial utilidad al marcar esta distinción. Segundo, la calidad del programa final de educación o tratamiento que sigue al diagnóstico fue muy importante. Tercero, la comprensión actual de la naturaleza de la inteligencia y sus determinantes surgió en repetidas ocasiones en los testimonios; nos referimos a algunas de estas cuestiones en el capítulo 7 al hablar de las teorías de la inteligencia. Aunque éstas parecen muy áridas, los desarrollos teóricos tienen implicaciones prácticas importantes en el mundo real, como en las salas de los tribunales. Por último, los métodos para analizar el sesgo de los reactivos usados en estos casos –podríamos decir, un análisis imaginario– parecen primitivos para los estándares actuales. Recordemos las descripciones en el capítulo 6 de los métodos contemporáneos para el estudio del sesgo de los reactivos.

Karraker contra Rent-A-Center

Una compañía usa el MMPI como una de varias medidas para seleccionar empleados para una promoción. Como describimos en el capítulo 13, el MMPI se usa, por lo general, para diagnosticar psicopatología; sin embargo, la compañía sostiene que usó la prueba como medida de personalidad, en particular para evaluar rasgos pertinentes para el desempeño en el puesto. ¿La prueba es un "examen médico" y, por ello, está sujeta a las disposiciones de Americans with Disabilities Act?

¿O es una prueba de rasgos pertinentes para el puesto y, por lo tanto, está sujeta a las disposiciones de "necesidad del negocio" de las Directrices EEOC?

En *Karraker contra Rent-A-Center, Inc.* (2005), "Karraker" se refiere, en realidad, a tres hermanos que trabajaban para Rent-A-Center (RAC) y solicitaron una promoción. Es interesante que el resumen del caso empieza refiriéndose al uso de pruebas, tanto de capacidad mental como de personalidad, con los candidatos a ser refuerzos en la *National Football League* (NFL). El razonamiento en este caso se volvió excepcionalmente complicado, con varios temas colaterales, pero al final el tribunal dictaminó que el MMPI era un examen médico, por lo que estaba sujeto a las disposiciones de ADA. En general, no puedes rechazar un empleo o promoción debido a una "discapacidad médica (o física)" y, de acuerdo con el tribunal, el MMPI estaba relacionado con dicho estatus de discapacidad.

Atkins contra Virginia

Daryl Atkins era un asesino convicto y sentenciado a muerte, sentencia que fue confirmada después de una apelación a la Suprema Corte de Virginia. Por medio del consejo municipal, Atkins apeló la sentencia a la Suprema Corte de EUA argumentando no su inocencia, sino que la sentencia, debido al retraso mental de Atkins, constituía "un castigo cruel e inusual" que violaba la Octava Enmienda de la Constitución de EUA. La

Suprema Corte de EUA, con una votación de 6 – 3, estuvo de acuerdo con el argumento y declaró que la sentencia de muerte no podía ser impuesta a una persona con retraso mental.[9]

Este caso tiene relevancia para el campo de las pruebas psicológicas debido a su papel para determinar el retraso mental, en especial las de capacidad mental general y de funcionamiento adaptativo, que se describieron en el capítulo 8. Las puntuaciones de las pruebas de CI tuvieron un papel decisivo en el caso *Atkins*; de modo que ellas, junto con otra información empleada para determinar el retraso mental, se vuelven pertinentes para cualquier futuro caso de pena capital.

El fallo *Atkins* de 2002 en esencia revocó el fallo de la Suprema Corte en un caso similar, *Penry contra Lynaugh* (1989), que concluyó que la imposición de sentencia de muerte no constituía automáticamente un castigo cruel e inusual para los retrasados mentales, aunque ese padecimiento podría ser considerado por el jurado para tomar una decisión. Después del fallo Atkins, la *American Psychiatric Association* (2002) publicó consejos para las legislaturas y los médicos sobre cómo aplicar el fallo. En general, los consejos reiteraban las definiciones profesionales de retraso mental (como las que presentamos en el capítulo 8) y hacían hincapié en la necesidad de la competencia profesional para aplicar e interpretar las pruebas.

Para el científico social, fue interesante cierta información incidental: el extenso uso de la Suprema Corte de datos de sondeos de opinión pública para llegar a un "consenso nacional" sobre qué tan apropiada es la pena de muerte para individuos con retraso mental. Los resúmenes de los resultados de 28 sondeos, compilados por la American Association on Mental Retardation, sirvieron como un apéndice a la opinión mayoritaria de la corte.

Moosman (2003) planteó una pregunta interesante acerca de qué otros "padecimientos psiquiátricos pueden ser tan discapacitantes, al menos, como el retraso mental" (p. 286), por ejemplo, los estados maniaco-depresivos, el TDAH o niveles bajos de serotonina en el cerebro pueden resultar en exenciones de la sentencia de muerte. Sólo el tiempo lo dirá; nos mantendremos al pendiente.

Caso Bomberos de New Haven

Su nombre oficial es *Ricci et al. contra DeStefano et al.* (2009), pero se conoce popularmente como caso Bomberos de New Haven. Este fallo incorporó nociones de la validez de las pruebas en relación con los requerimientos del título VII de la Ley de Derechos Civiles, que ya hemos descrito. La ciudad de New Haven, Connecticut, usó una prueba para seleccionar bomberos para una promoción. En la aplicación de 2003 de la prueba, siguiendo las reglas establecidas por los procedimientos del servicio civil de la ciudad, 10 candidatos blancos se volvieron elegibles para la promoción inmediata para teniente, mientras que siete blancos y dos hispanos cumplieron con los requisitos para ser promovidos a capitán. Ningún candidato negro fue digno de promoción. El Consejo de la Ciudad se negó a avalar los resultados debido a la disparidad racial. Los bomberos, cuya

promoción fue bloqueada, presentaron una demanda afirmando que los resultados debían ser avalados. (Ricci era uno de los bomberos y DeStefano era el alcalde de la ciudad). En un fallo cerrado de 5 – 4, la Suprema Corte de EUA revocó la resolución del tribunal de menor jerarquía y encontró que el rechazo de los resultados de las pruebas por parte del Consejo de la ciudad fue, en esencia, un caso de discriminación inversa.

El argumento clave en el caso se centró en la validez de la prueba de selección. El fallo de la corte citó el testimonio acerca del proceso de elaboración de la prueba, incluyendo los esfuerzos para establecer su pertinencia para el trabajo de los reactivos (véase capítulo 5 sobre la validez de las pruebas de empleo). La corte concluyó que "la afirmación de la ciudad de que los exámenes en cuestión no estaban relacionados con el puesto ni eran consistentes con la necesidad del negocio fue contradicha descaradamente por los documentos".

Las primeras audiencias de este caso se centraron en la validez de la prueba de selección; sin embargo, conforme avanzaron las deliberaciones, la atención se dirigió cada vez más hacia los motivos del Consejo de la Ciudad para rechazar los resultados de las pruebas. Al final, la corte concluyó que el consejo temía un litigio y disturbios civiles por aceptar los resultados, pero la corte concluyó que ese temor no justificaba la violación de derechos de los bomberos que habían cumplido con los requisitos para la promoción.

El expediente del fallo New Haven revela dos datos incidentales interesantes. Primero, en una de las opiniones discrepantes, un juez de la Suprema Corte usó de manera consistente el término "confiabilidad" cuando era evidente que se refería a la validez de la prueba. Segundo, uno de los asesores empleados por la Junta de Servicios Civiles recomendó ajustar los resultados de las pruebas para asegurar la promoción a cierto número de candidatos de las minorías. Hacer esto habría violado sin duda las disposiciones de la Ley de Derechos Civiles de 1991, que prohíben "la obtención de normas por subgrupos", como se describió en este capítulo.

Aplicaciones forenses de las pruebas «423-424a

La **psicología forense** es una especialidad que surgió con rapidez. Por lo general se ocupa de la aplicación de los principios y métodos psicológicos en el contexto legal, en especial, pero no exclusivamente, en las acciones en los procesos judiciales. Algunos psicólogos se especializan en este campo; nosotros nos concentramos aquí en cualquier aplicación de las pruebas dentro del contexto legal. Por ejemplo, un psicólogo clínico puede utilizar pruebas cuando funge como testigo experto en la competencia mental de un acusado; un neuropsicólogo puede testificar para documentar el grado del daño cerebral en un caso de lesión personal; o un psicómetra puede testificar acerca de la calidad técnica de una prueba empleada para identificar el posible retraso mental de un niño. Se puede consultar Otto y Weiner (2013) para conocer el estatus actual de la especialización forense.

Las aplicaciones forenses de las pruebas se dividen en tres categorías importantes. Primero, hay dos términos legales tradicionales con un significado claramente psicológico que pueden exigir el uso de pruebas. Segundo, hay tres áreas de importancia especial para las aplicaciones forenses. Tercero, hay un amplio rango de áreas en que los psicólogos aplican los métodos de evaluación a las actividades de los tribunales.

Dos términos legales

Dos términos legales tienen un toque claramente psicológico: demencia y competencia para comparecer ante un juicio. Estos términos tienen una larga historia en leyes comunes en EUA e Inglaterra. Para los novatos, pueden sonar como la misma cosa, pero tienen significados muy diferentes para la ley. **Demencia** se refiere al trastorno mental o incapacidad *en el momento en que se comete un crimen*. **Competencia para comparecer ante un juicio** se refiere a la capacidad mental de una persona en el momento del juicio por un crimen.[10] La distinción esencial entre los términos, aunque no es la única, es la temporalidad.

Históricamente, ha habido tres reglas para definir la demencia, cuyos nombres se originaron con los casos en que el tribunal las enunció. La regla M'Naghten dice que la persona no puede distinguir entre el bien y el mal; se trata de un criterio cognitivo. La regla Durham, que es más liberal, dice que la persona cometió el acto a causa de un trastorno mental; se trata de un criterio volitivo. La regla Brawner, también conocida como la Regla del Instituto de Leyes de EUA, admite la incapacidad para distinguir el bien del mal o para controlar la conducta, exclusiva de ciertos padecimientos como la psicopatía o las compulsiones.

En 1982, John Hinckley, quien había intentado asesinar al presidente Reagan, fue absuelto porque la defensa argumentó demencia. El Congreso de EUA, con un disgusto

claro por el resultado, aprobó la Ley de Reforma a la Defensa por Demencia. Esta ley limitó la defensa por demencia a las pruebas cognitivas y estableció un estándar alto para demostrarla. Además, una persona encontrada demente en un proceso criminal tenía que ser remitida a una institución mental.

La competencia para comparecer ante un juicio implica la capacidad mental de un individuo para comprender los procesos del juicio y trabajar razonablemente con asesoría legal. Si hay una declaración de incompetencia, debe haber una audiencia sobre ella antes del juicio por el presunto crimen. Por lo común, una persona encontrada incompetente sería remitida a una institución mental; sin embargo, si después se encuentra competente, la persona puede ser juzgada por el presunto crimen.

Es evidente que los resultados de las pruebas psicológicas pueden ser muy importantes para determinar una demencia o la competencia para comparecer ante un juicio. Ya que el alegato de demencia se relaciona con la capacidad mental en el momento del crimen, que puede ser meses o incluso años antes del juicio, los resultados de pruebas disponibles previamente podrían ser muy pertinentes. Las pruebas aplicadas cerca del momento del juicio pueden no ser muy útiles para juzgar el estado del acusado al momento del crimen, pero sí para determinar la competencia para comparecer ante un juicio.

Tres áreas de especial interés

Más allá del tema de la demencia y la competencia para comparecer ante un juicio, en años recientes, los psicólogos han encontrado que las pruebas son potencialmente útiles en tres áreas forenses. Primero, la evaluación puede ser útil en *casos relacionados con la custodia de un niño*. Aquí, hay un intento de determinar los requisitos de los padres (o de otras partes) para servir de la mejor manera a los intereses del niño. Segundo, la metodología de la evaluación se ha aplicado a la **predicción de conductas violentas o de riesgo,** lo que a veces se denomina peligrosidad. Estas predicciones pueden ser importantes para determinar, por ejemplo, a quién se puede conceder libertad condicional, o a quién se debe encarcelar o ser puesto en libertad bajo supervisión. Tercero, la evaluación psicológica puede ser útil para determinar la **naturaleza y el grado del abuso,** en casos de abuso marital o sexual. Estas tres áreas aún necesitan desarrollos sustanciales.

Y más allá

Más allá de las áreas mencionadas, las pruebas pueden formar parte de la actividad forense de maneras prácticamente ilimitadas. Cualquier uso de una prueba psicológica tiene el potencial de formar parte de los procesos judiciales. Debido a la predilección estadounidense para litigar por todo, el uso forense de las pruebas, sin duda, seguirá creciendo.

En general, los tipos de pruebas que hemos revisado en este libro dan una buena

indicación de las pruebas que pueden aparecer en la sala de un tribunal. Lees-Haley (Lees-Haley, Smith, Williams, & Dunn, 1996) estudió los tipos de pruebas usadas por los psicólogos forenses, y sus resultados, por lo general, concuerdan con los de otros estudios. Las pruebas usadas en el contexto forense son las mismas que las de otros contextos: WAIS, MMPI, WMS, Rorschach, etc.

Algunas generalizaciones acerca de la conexión del campo de las pruebas y la ley

Al considerar la conexión del campo de las pruebas y la ley, desarrollamos las siguientes generalizaciones (véase Resumen de puntos clave 16-4). Primero, todo lo que hemos revisado confirma la importancia de la calidad técnica de las pruebas. Los conceptos de validez, confiabilidad y adecuación de las normas ya no son sólo para los profesionales expertos en psicometría, pues ahora están escritos en las leyes y son puntos de referencia en los procesos judiciales. Podemos señalar ejemplos en los que una ley o el fallo de un tribunal está en conflicto con una práctica psicométrica adecuada o no la toma en cuenta, pero éstos tienden a ser las excepciones. La regla general es: la ley apoya y, además, depende de la buena psicometría. Además, la competencia técnica del usuario de la prueba es esencial. La ley demanda tal competencia, por lo que los tribunales esperan un alto nivel de competencia. Esta generalización concuerda muy bien con el requerimiento ético de competencia discutido en este capítulo.

Segundo, los requerimientos legales que evolucionaron en los últimos 50 años han influido en la elaboración y el uso de las pruebas. La influencia más importante ha sido el acrecentado interés en la aplicabilidad de las pruebas a varios subgrupos. Hay un particular interés en los subgrupos raciales/étnicos y en las personas con discapacidades. Las influencias son más evidentes respecto de a) la elaboración del contenido de las pruebas, b) la estandarización y c) los estudios de validez. En cuanto al contenido de las pruebas, en el capítulo 6 describimos los métodos para detectar sesgos recurriendo a paneles de revisión y a estudios de funcionamiento diferencial de los reactivos (FDR). Ahora, aplicar estos métodos es una práctica de rutina.

En cuanto a la elaboración de normas, ahora las editoriales de pruebas intentan representar de manera cuidadosa a diversos subgrupos en programas de estandarización, en lo cual han hecho grandes avances. Es importante señalar que la falta de una representación adecuada de varios subgrupos en las normas puede (o no) afectar en la exactitud de las normas, pero esto no tiene nada que ver con la validez o la confiabilidad de la prueba. Esta generalización parece ir en contra del sentido común y los legos nunca la creen. Para un psicómetra novato, una prueba que no tiene, digamos, ningún afroamericano en el grupo de estandarización es inválida automáticamente para aplicarla a afroamericanos. Esto no es cierto; sin embargo, los sentimientos en este punto están tan profundamente arraigados que no vale la pena discutir sobre este punto fuera de los círculos de la psicometría. A la inversa, una norma con representación impecable de varios grupos no garantiza la validez o confiabilidad de la prueba al aplicarse a esos grupos.

En cuanto a la validez, es claro que el ambiente legal ha forzado a los autores y usuarios de pruebas a demostrar que éstas son válidas para grupos específicos. Por ejemplo, ¿el WISC es válido para niños con dislexia? ¿El SAT es válido para estudiantes

hispanos? Este énfasis ha producido una gran cantidad de estudios acerca de cómo funcionan las pruebas con grupos específicos: un desarrollo muy saludable.

Tercero, nuestra revisión de los procesos judiciales nos recuerda que no debemos sobregeneralizar a partir de casos únicos. Los fallos no siempre son consistentes, ni por completo predecibles. Quizá más importante, un fallo a veces tiene un ámbito muy restringido o depende de un razonamiento relacionado con circunstancias especiales. Véase, por ejemplo, el fallo del caso *Larry P.* que describimos antes. El fallo del caso *Debra P.* dependió, en parte, de estudios muy específicos de validez instruccional realizados en el estado de Florida. Aunque nuestro sistema legal, por lo general, opera con base en los precedentes, éstos no siempre son claros e inequívocos, y a veces no se aplican en absoluto. Este asunto de la conexión entre el campo de las pruebas y la ley no es sencillo; las leyes y regulaciones pueden tener un propósito general claro, pero sus detalles pueden dejarnos helados. Es comprensible que los procesos judiciales estén repletos de una jerga casi imposible de comprender, por no decir imposible; sin embargo, presentan retos especiales más allá de conocer los métodos y principios de las pruebas psicológicas.

Resumen de puntos clave 16-4

Generalizaciones acerca de la conexión entre el campo de las pruebas y la ley

La ley confirma la importancia de la calidad técnica de las pruebas y la competencia del evaluador.

La ley ha influido de manera importante en la elaboración y uso de pruebas.

La conexión es compleja, pues requiere conocimiento especializado y exige precaución al hacer generalizaciones a partir de casos únicos.

Resumen

1. La ética se ocupa de lo que se debe o no se debe hacer, de acuerdo con principios éticos, morales y profesionales. La ley se ocupa de la obligación de hacer o de no hacer de acuerdo con las leyes.
2. Un código de ética profesional ayuda a definir qué significa "hacer el bien" en el contexto de los temas que con frecuencia se encuentran en una profesión. Así, el código sensibiliza al profesional y establece también las expectativas sociales, con lo cual protege la reputación de la profesión.
3. Las raíces lejanas de los códigos éticos relacionadas con las pruebas se encuentran en el juramento hipocrático, el código Nuremberg y el Informe Belmont. Las raíces más inmediatas se encuentran en los *Standards for Educational and Psychological Tests* (AERA, APA, NCME, 2013) y en el Código ético de la APA. Otras asociaciones profesionales tienen códigos éticos similares.
4. Identificamos cinco principios éticos con aplicaciones muy amplias en el campo de las pruebas: asegurar la competencia, obtener el consentimiento informado, proporcionar el conocimiento de los resultados, mantener la confidencialidad y resguardar la seguridad de las pruebas. También identificamos tres principios con una aplicabilidad más limitada: fijar estándares altos para la elaboración/publicación de pruebas, asumir la responsabilidad por los informes de pruebas hechos de manera automatizada y trabajar para prevenir el uso no calificado de pruebas.
5. Muchas editoriales de pruebas emplean un sistema de tres niveles (A, B y C) para clasificar las pruebas y a los potenciales compradores.
6. Hay tres tipos de "leyes": legislación, regulaciones administrativas y jurisprudencia.
7. Las leyes afectan el campo de las pruebas sobre todo en estas áreas de aplicación: empleo, educación, programas autorizados de evaluación y psicología forense.
8. Las principales leyes que afectan el uso de las pruebas son:

- Decimocuarta Enmienda a la Constitución de EUA
- Ley de Derechos Civiles de 1964 y 1991
- Rehabilitation *Act y Americans with Disabilities Act* (ADA)
- Varias leyes relacionadas con la educación de personas con discapacidades, en especial IDEA, FERPA e HIPPA
- Las Directrices EEOC sobre las pruebas para dar empleo, en especial respecto de los requerimientos de información de validez y la definición del impacto adverso
- *No Child Left Behind Act*

9. Varios procesos judiciales ilustraron lo que los tribunales esperan de las pruebas. En el caso *Giggs*, fue crucial establecer la validez de los requerimientos del empleo; en los de *Debra P.* y *GI Forum*, era aceptable una prueba autorizada por el estado para

otorgar el certificado de bachillerato siempre y cuando la prueba tuviera validez instruccional y las escuelas proporcionaran la preparación adecuada. Los casos *Larry P.*, *PASE* y *Crawford* produjeron fallos confusos y en conflicto entre sí acerca del uso de las pruebas individuales de inteligencia. El caso *Karraker* mostró que el modo en que una prueba se usa de manera regular influye en su aceptabilidad legal. El caso *Atkins* dio por resultado la prohibición de la pena de muerte basada, en gran parte, en los resultados de pruebas de capacidad mental. En el caso *New Haven*, la intención de usar (o no usar) los resultados de las pruebas fue importante.

10. Dentro de la especialidad de psicología forense, las pruebas pueden ser muy útiles con respecto a dos conceptos legales bien establecidos: demencia y capacidad para comparecer ante un juicio. Las pruebas también pueden ser útiles en casos relacionados con la custodia de un niño, con la predicción de conducta violenta futura y con abuso. Más allá de estas aplicaciones específicas, el rango completo de usos normales de las pruebas tiene el potencial de intervenir en procesos legales.

11. Respecto de la conexión entre el campo de las pruebas y la ley, desarrollamos generalizaciones acerca de:

- La importancia de la calidad técnica y la competencia
- La influencia de las leyes en la elaboración y uso de las pruebas
- La complejidad de la conexión entre el campo de las pruebas y la ley y la necesidad de tener cuidado al formular conclusiones a partir de casos únicos

Palabras clave

C.F.R.
Código ético de la APA
competencia
competencia para comparecer ante un juicio
confidencialidad
conocimiento de los resultados
consentimiento informado
demencia
Leyes
ADA
Decimocuarta Enmienda
Directrices EEOC
Procesos judiciales
Atkins
Crawford
Debra P.
disposición del debido proceso
disposición de protección igualitaria
ética
IEP
impacto adverso
ley
ley administrativa o regulaciones
ley estatutaria o legislación
FERPA
HIPPA
IDEA
Ley de Derechos Civiles
GI Forum
Griggs
Karraker
Larry P.
obtención de normas por subgrupos
P.L.
proceso judicial
psicología forense
regla de cuatro quintos
requisitos para el usuario de pruebas
seguridad de las pruebas

U.S.C.
No Child Left Behind
Rehabilitation Act
New Haven
PASE

Ejercicios

1. Regresa a los casos de la página 404a» y consulta el resumen de los principios éticos relacionados con el campo de las pruebas de la página 410a». ¿Qué principios se aplican a cada caso?

2. Señalamos que varias asociaciones profesionales tienen códigos de ética relacionados con el campo de las pruebas y que son muy similares. Puedes consultar en línea los códigos de la *American Psychological Association, American Counseling Association y National Association of School Psychologists.* Entra al menos a dos de estos códigos y revisa sus semejanzas. Aquí están las direcciones: www.apa.org; www.counseling.org; www.nasponline.org

4. Entra a las páginas de internet de dos editoriales de las que se enumeran en el apéndice C o consulta la versión impresa de sus catálogos. ¿Cuáles son los requisitos que deben cubrir los usuarios de pruebas de acuerdo con cada editorial? ¿Emplean el sistema de tres niveles esbozado en la página 409a» de este libro?

5. Averigua si la biblioteca de tu universidad está suscrita a LexisNexis o Westlaw. Si es así, escribe las palabras clave de uno de los procesos judiciales que presentamos, por ejemplo, *Atkins contra Virgina*, *Griggs contra Duke Power* o *Debra P. contra Turlington*. Observa cómo se enmarca el tema al principio del proceso y cómo se resume el fallo al final.

6. Ve al ejemplo para determinar el impacto adverso en el cuadro 16-9.Cambia los números de la siguiente manera: blancos – 26, negros – 13, asiáticos – 4, hispanos – 7. Calcula las nuevas tasas de selección y "TS/tasa más alta". ¿Hay impacto adverso en algún grupo?

7. El texto completo de las leyes IDEA y *No Child Left Behind* puede encontrarse en cualquier página general sobre leyes federales. Sin embargo, estas dos tienen sus propias páginas de internet, las cuales contienen leyes, regulaciones asociadas y otros materiales. Si te interesan estas leyes, revisa el sitio del *Office of Special Education and Rehabilitative Services* [Ministerio de Educación Especial y Servicios de Rehabilitación] del *Department of Education* [Departamento de Educación] de EUA: www.ed.gov

---

Notas

[1] El gobierno de EUA juega continuamente con la terminología exacta para estos grupos. Para confirmar las denominaciones y definiciones de inclusión más recientes, haz una búsqueda en internet sobre la "clasificación del gobierno de EUA por grupo minoritario".

[2] La Quinta Enmienda también tiene una cláusula del "debido proceso", que tiene que ver con asuntos como doble peligro, autoincriminación y otras cuestiones que no son pertinentes para nuestras consideraciones.

[3] El título oficial de esta revisión es Ley de Mejoras a la Educación de Individuos con Discapacidades de 2004, del que surgió el acrónimo IDEIA, aunque suele citarse con mayor frecuencia como IDEA 2004.

4 En este ejemplo, también damos por hecho una variabilidad equivalente en los subgrupos.

5 Cuando citamos procesos judiciales en esta sección, por lo general, citamos la disposición final del caso y no incluimos todas las formulaciones originales, apelaciones y arrestos preventivos. Citar la disposición final llevará al lector a las primeras etapas del proceso.

6 Al igual que en el caso Debra P., que discutimos en la siguiente sección, el pleito en realidad usó tres subclases, pero no son importantes para nuestro resumen.

7 Conforme avanzó el caso, se formaron tres clases, pero este punto no es importante para nuestro resumen. Numerosos funcionarios, además de Turlington, fueron nombrados como acusados, pero este hecho tampoco es importante aquí.

8 El registro del tribunal emplea una declaración errónea de la regla, que define "impacto adverso (cuando) la tasa de aprobación de un grupo minoritario es menor del 80% de la tasa de aprobación del grupo mayoritario". Véase la p. 591 para conocer la verdadera regla EEOC de manera literal.

9 Eventos subsiguientes volvieron fascinante el caso Atkins, pero son irrelevantes para el punto básico: el retraso mental impide la pena capital.

10 La competencia también se aplica en la ley civil, por ejemplo, para la preparación de un testamento. Aquí sólo tratamos la aplicación en la ley criminal.

APÉNDICE A
Revisión y selección de pruebas

Este apéndice ayudará al estudiante a realizar proyectos de revisión o selección de pruebas como ejercicios de entrenamiento apegados a la *Declaración sobre el uso de pruebas psicológicas seguras en la educación de estudiantes de licenciatura y posgrado en psicología* de la American Psychological Association. En la página http://www.apa.org/science/leadership/tests/test-security.aspx se puede encontrar una copia de esta declaración.

Hay dos tareas prácticas paralelas a los dos temas introducidos al principio del capítulo 2 que los psicólogos llevan a cabo. La primera es revisar una prueba, por ejemplo, el ABC Personality Inventory o el Pennsylvania Non-verbal Intelligence Test. La segunda tarea es conjuntar una serie de pruebas cuyo uso pueda contemplarse para un propósito específico. Por ejemplo, ¿qué pruebas deben considerarse para el programa diagnóstico de lectura del distrito escolar o para la exploración inicial de clientes en una clínica? No hay reglas rígidas específicas para responder a este tipo de preguntas; en algunas circunstancias, las soluciones a estos problemas podrán encontrarse con rapidez después de una breve consulta en una de las fuentes de información presentadas en el capítulo 2. En otras circunstancias se formularán más preguntas formales. Este apéndice esboza las maneras típicas de responder a estas preguntas.

Estándares absoluto y relativo

Antes de bosquejar los procesos para revisar y seleccionar pruebas, puede ser útil contrastar dos métodos de estas actividades, que se relacionan con el tipo de estándares empleados para presentar los juicios sobre las pruebas. Aquí nos referimos a ellos como el estándar absoluto y el estándar relativo. Diferentes revisiones adoptan uno u otro estándar, a menudo sin identificar explícitamente de cuál de los dos se trata ni insinuar que alguno de ellos está implicado.

El *estándar absoluto* emplea los niveles más altos de excelencia como punto de referencia para la construcción de pruebas, validación, normalización, etc. De acuerdo con este estándar, por ejemplo, la confiabilidad de una prueba debe determinarse mediante diversos métodos para distintos grupos de examinados con coeficientes de confiabilidad igualmente altos. Debe haber una gran cantidad de estudios de validez disponibles sobre el instrumento. Debe haber disponibles normas nacionales basadas en muestras grandes, elegidas con cuidado, además de normas para varios tipos de subgrupos, y lo mismo debe ocurrir con otras características de las pruebas. Cualquier

prueba que no alcance estos altos niveles de excelencia será criticada con dureza y su uso no se recomendará. A algunas pruebas les fue bien cuando fueron juzgadas de acuerdo con el estándar absoluto.

De acuerdo con el *estándar relativo*, el valor de una prueba se juzga primordialmente en relación con otras pruebas del mismo dominio o con no usar ninguna. ¿Esta prueba sirve mejor que otras a los propósitos del usuario en vista de que, a pesar de sus defectos, es mejor que las demás? ¿O para los propósitos del usuario es mejor usar esta prueba, tan defectuosa como pueda ser, que no usar ninguna?

En el trabajo aplicado, el estándar relativo, por lo general, prevalece sobre el absoluto. Uno va a elegir una prueba, sea ésta o aquélla. O uno va a decidir que ese no usar ninguna prueba es un mejor procedimiento que usar alguna prueba. Cualquiera de estas decisiones se basa en una comparación de varias alternativas. Por otro lado, el estándar absoluto se usa a menudo en reseñas de pruebas y en libros de texto para presentar algunas pruebas. Suele ser más difícil aplicar el estándar relativo, pues se necesita saber no sólo acerca de la prueba, sino también de otras posibles alternativas.

Reseñas de pruebas

Las tradiciones para realizar la reseña de una prueba están establecidas por las prácticas del *Mental Measurements Yearbooks* (MMY) de Buros y la serie de *Test Critiques* descrita en el capítulo 2. Aunque estas dos publicaciones han sugerido lineamientos algo diferentes para sus reseñas, éstas se organizan de manera similar. Las sugerencias que damos aquí se inspiran en los formatos de estas dos fuentes clásicas de reseñas de pruebas. El cuadro A-1 resume los pasos para preparar la reseña de una prueba, y el cuadro A-2 identifica sus principales partes.

Cuadro A-1. Pasos para disponerse a hacer una reseña

1. Reúne los materiales para la reseña.
2. Determina el propósito de la prueba.
3. Examina los materiales de la prueba.
4. Revisa las características técnicas de la prueba:
  a. Confiabilidad
  b. Validez
  c. Tipos de normas y proceso de normalización
  d. Elaboración de la prueba

Las reseñas del MMY no siguen un formato fijo. Los autores ejercen cierta libertad en la organización y el uso de los subtítulos. Para ver las recomendaciones del Instituto Buros para los autores de reseñas, visita http://buros.org/reviewers. Sin embargo, todas las reseñas cubren los temas que se esbozan en el cuadro A-2. Por otro lado, las reseñas de *Test Critiques* siguen un formato fijo. Las reseñas del MMY están precedidas de un listado sencillo de los materiales de la prueba, servicios de calificación, etc., que

proporcionan los editores del MMY, no el autor de la reseña.

Cuadro A-2. Partes principales de una reseña
1. Formula el propósito de la prueba. Describe los fundamentos del método adoptado. **2.** Describe la estructura y los materiales de la prueba: tipos de reactivos, niveles, extensión, etc. **3.** Por cada una de las siguientes áreas, indica de qué información técnica se dispone y haz un juicio con base en ella. Estos temas pueden abordarse en cualquier orden. **a.** Confiabilidad **b.** Validez **c.** Normas **d.** Procedimientos para elaborar la prueba **4.** Presenta un resumen acerca de la calidad de la prueba y su utilidad potencial.

Materiales necesarios

El primer paso para realizar una reseña es reunir los materiales necesarios. Si es acerca de una prueba publicada, se requieren dos tipos de materiales. Primero, una prueba publicada incluye materiales como el manual, cuadernillos o estímulos, hojas de respuesta u otros dispositivos para registrar las respuestas, claves de calificación, instrucciones y materiales relacionados. En el caso de pruebas complejas como las baterías de aprovechamiento de niveles múltiples, se incluye sólo una selección de los materiales; en estos casos, el autor de la reseña necesitará ponerse en contacto con la editorial para obtener un conjunto más completo de los materiales. La prueba también debe incluir un manual, en el que se expone con claridad su propósito y fundamentos, las instrucciones para su aplicación, cuadros de normas e información sobre confiabilidad, validez, procedimientos de elaboración y programas de normalización de la prueba. En el caso de algunas pruebas, estos temas pueden estar divididos en dos o más manuales; por ejemplo, un manual puede contener las instrucciones de aplicación y calificación, mientras que otro, la información psicométrica.

El segundo tipo de información importante en una reseña es la entrada acerca de la prueba del catálogo de la editorial. (Evidentemente, esto se aplica sólo a las pruebas publicadas.) El catálogo describe la variedad de los materiales, por ejemplo, el número de niveles y formas, el tipo de servicios de calificación y los costos de los materiales de la prueba. Gran parte de esta información está disponible en la página de internet de la editorial. En el apéndice C, encontrarás una lista de las páginas de las principales editoriales de pruebas.

La reseña se concentra en el examen de los materiales y las características de la prueba, pero antes de empezar no es raro, aunque no siempre es necesario, ponerse en contacto con el autor o la editorial para asegurarse de que se dispone de los materiales completos y actuales para el proceso de escribir la reseña.

Hay varios pasos que pueden ser aceptables como parte del proceso de escribir la reseña, pero que, de hecho, *no* suelen llevarse a cabo. Primero, el autor de la reseña, por

lo general, *no* hace una búsqueda meticulosa de la literatura acerca de la prueba. Los materiales deben tener un espacio propio para propósitos de la reseña. El lector puede consultar otras reseñas de la prueba, en especial de ediciones anteriores (si hay alguna) o algunos hallazgos de la investigación acerca de la prueba. Sin embargo, no es habitual realizar una búsqueda exhaustiva de la literatura cuando se está reseñando una prueba.

Segundo, cuando se reseña una prueba, *no* es común llevar a cabo una prueba piloto extensa, por ejemplo, aplicarla a docenas o incluso a cientos de individuos. No sería raro simular una aplicación regular a uno o dos individuos; sin duda, el autor de la reseña simularía que contesta la prueba completa o una parte. Sin embargo, la aplicación piloto extensa no es habitual. Por otro lado, si uno considera adoptar realmente la prueba para usarla de manera regular, es muy recomendable realizar alguna aplicación piloto.

Estructura de la reseña

Como señalamos, la reseña de una prueba se concentra en el examen de los materiales y las características técnicas de la prueba. El cuadro A-2 sirve como guía de la estructura de una reseña. El primer paso decisivo es determinar el propósito de la prueba, que debe estar formulado con claridad en el manual. También puede aparecer en el catálogo de la editorial.

Después, el autor debe examinar de manera meticulosa los materiales. ¿Qué es esta prueba con exactitud? ¿Qué reactivos u otros estímulos tiene? ¿Qué hace el examinado? ¿Cuánto tiempo requiere su aplicación? Y así sucesivamente. Para responder este tipo de preguntas, es común que el autor de la reseña "simule" que contesta realmente la prueba. El autor también lee con detenimiento las instrucciones, los procedimientos de calificación y, además, se mete de lleno en los materiales de la prueba. También es importante determinar qué puntuaciones produce la prueba.

Por último, el autor empieza el examen de las características técnicas de la prueba. Para los novatos, ésta es sin duda la parte más abrumadora de esta tarea. Sin embargo, una solución, por lo general, es plantearse las siguientes cuatro preguntas:

1. ¿Qué evidencia existe de la confiabilidad de las puntuaciones de la prueba?
2. ¿Qué evidencia existe de la validez de las puntuaciones o usos de la prueba?
3. ¿Qué evidencia existe de la utilidad de las normas de la prueba?
4. ¿Qué procedimientos se usaron para elaborar la prueba y asegurar que se cumpliría su propósito?

Estas preguntas corresponden a los capítulos 3-6 de este libro; de modo que responder estas preguntas es un ejercicio en el que se aplica lo aprendido en esos capítulos.

Después de conocer minuciosamente los materiales y las características técnicas de la prueba, el autor está listo para escribir la reseña. La extensión normal y el estilo de las reseñas pueden determinarse consultando cualquier edición del *MMY* o de *Test Critiques*.

Las reseñas del *MMY* van de 500 a 2000 palabras. (En http://buros.org/review-samples se pueden ver ejemplos.) Por su parte, las reseñas de *Test Critiques* tienden a ser un

poco más extensas; la mayoría va de 2000 a 2500 palabras, pero algunas llegan a tener 6000.

La reseña empieza con una sección que es casi por completo descriptiva y sin juicios. La oración inicial identifica el propósito de la prueba y, a menudo, se toma directamente del manual. Por ejemplo:

> De acuerdo con el manual de la prueba (p. 2), el ABC-PI busca ofrecer una medida sencilla de los principales rasgos de personalidad del adolescente normal.

La reseña puede elaborar de manera breve el propósito parafraseando secciones del manual, por ejemplo:

> Los principales rasgos de personalidad que cubre esta prueba incluyen...

La población a la que está dirigida la prueba suele identificarse aquí. En esta sección inicial también se discute cualquier orientación teórica que pueda tener la prueba. Por ejemplo:

> La estructura de la prueba intenta operacionalizar la teoría de las cuatro etapas del desarrollo adolescente de Smith.

En esta sección inicial, el autor no hace comentarios críticos sobre el propósito de la prueba. Sin embargo, puede comentar sobre qué tan clara es la formulación del propósito.

Después de identificar el propósito y los fundamentos de la prueba, la reseña continúa con la descripción de la estructura de la prueba y sus materiales. Esta sección puede considerarse como una fotografía verbal de la prueba. ¿Cuáles son los materiales de la prueba? ¿Cuántos reactivos la integran? ¿De qué manera debe responder el examinado? ¿Existen diferentes niveles y formas de la prueba? Y así sucesivamente. En esta sección se supone que el lector de la reseña no tiene los materiales de la prueba, por lo que es tarea del autor dar una especie de visita guiada por la prueba.

Por último, la sección inicial dedicará espacio a cualquier versión especial o características poco comunes de la prueba, si las hay. Por ejemplo, si la prueba está disponible en español e inglés o si hay una prueba opcional de práctica, estas cuestiones deben señalarse con brevedad.

En la segunda sección importante de la reseña se tratan las características técnicas de la prueba combinando la descripción con el juicio. Es conveniente usar cuatro subtítulos, como se bosqueja en los incisos 3a-3d del cuadro A-2. En cada una de estas partes, el autor de la reseña describe la evidencia disponible de la prueba, y luego presenta sus juicios acerca de qué tan adecuada es esta información. Por ejemplo, respecto de la confiabilidad, la reseña puede señalar que:

> Los coeficientes alpha de las tres puntuaciones de la prueba, con base en las muestras de normalización, son .63, .85 y .91. La primera puntuación es claramente inadecuada para usarla de manera habitual, pero si se usa, se deben extremar las precauciones. Las otras dos muestran una consistencia interna adecuada para usarlas de manera normal. Sin embargo, el manual de la prueba no ofrece información sobre la confiabilidad de test-retest, lo cual constituye un defecto importante.

La reseña ofrece un tratamiento similar de las otras características técnicas. El patrón siempre es: una descripción seguida de un juicio.

La sección final de la reseña ofrece una evaluación general de la prueba. El autor puede adoptar una de varias estrategias posibles en esta sección. Primero, los comentarios pueden limitarse a una sinopsis de las principales conclusiones de las primeras secciones de la reseña. Segundo, el autor puede sugerir cautela para tener en cuenta si la prueba se usa y/o hacer recomendaciones para el autor de la prueba o la editorial para que las sigan en el posterior desarrollo de la prueba. Estas precauciones y recomendaciones serán resultado de las primeras partes de la reseña. La tercera estrategia, la más audaz, hace una recomendación específica en favor o en contra del uso de la prueba, a veces, con una referencia a otra prueba. El autor de la reseña puede recomendar esta prueba por encima de otras alternativas diciendo, por ejemplo, "Para los clínicos interesados en una evaluación inicial rápida de la capacidad mental, la Prueba de Inteligencia de Hogan es, con claridad, la primera opción". O puede recomendar no usar la prueba diciendo, por ejemplo, "Hay varias pruebas bien conocidas que serán preferibles a la Prueba de Inteligencia de Hogan". Desde luego, el autor de la reseña ofrece los fundamentos de su recomendación recurriendo a los comentarios de las primeras secciones de la revista. Es importante que la conclusión sea consistente con las primeras secciones.

Selección de la prueba

El proceso para seleccionar una prueba está diseñado para encontrar una que cumpla de manera adecuada algún propósito. Este proceso es similar en ciertos aspectos al de reseñar una prueba, pero en otros aspectos son dos procesos muy diferentes. La diferencia clave es el punto de inicio; en el caso de la reseña, empieza con una prueba específica, mientras que en el de selección, con un propósito particular y luego continúa buscando hasta encontrar la prueba que sirva a ese propósito. Una diferencia curiosa incidental entre los dos procesos es que la reseña suele hacerla un individuo, mientras que, en la práctica, la selección de una prueba, a menudo, es realizada por un comité.

El cuadro A-3 resume los cinco pasos principales del proceso de selección de una prueba. El primer paso es definir con claridad el propósito al que servirá la prueba. ¿Por qué se necesita la prueba? ¿Quién usará las puntuaciones? A primera vista, este paso parece sencillo; en algunos casos, así es. Por ejemplo, el solo propósito de que la prueba sirva como variable dependiente en un pequeño proyecto de investigación. Sin embargo, en otros casos, puede haber muchos usos de las puntuaciones de la prueba; por ejemplo, las de una prueba de aprovechamiento podrían ser usadas por los maestros para identificar a estudiantes con una necesidad de ayuda especial, por el consejo escolar local para un informe al público, por el psicólogo escolar para ayudar a definir dificultades de aprendizaje o por los padres para monitorear el progreso de sus hijos. Una prueba de personalidad podría usarse en una clínica, en parte, para proveer una base para la discusión inicial con clientes y, en parte, como medida previa de un proyecto de

investigación en marcha que abarca una red de clínicas.

Cuadro A-3. Pasos del proceso de selección de una prueba
1. Definir con claridad el propósito y el uso al que está destinada, el grupo meta y las restricciones.
2. Compilar una lista inicial de candidatas.
3. Compilar una lista de finalistas.
4. Resumir la información de las finalistas.
5. Hacer una selección o recomendación.

Por lo tanto, se debe dedicar el tiempo suficiente a este paso inicial para definir con claridad el propósito de la prueba. Los usos anticipados de las puntuaciones merecen atención especial, pues la reflexión puede revelar un propósito primario y varios secundarios.

También es importante especificar en esta etapa a) cualquier condición especial para la aplicación de la prueba y b) las características de la población meta. Algunas condiciones pueden ser esenciales y otras, sólo deseables; por ejemplo, puede ser esencial que la prueba esté disponible en lenguaje Braille, y puede ser deseable, pero no esencial, que su aplicación requiera menos de 50 min. Las características pertinentes de la población meta incluyen edad, nivel educativo o capacidad mental, lengua materna y posibles discapacidades. Por ejemplo, la población meta pueden ser individuos con nivel limítrofe de alfabetismo, por lo que se requiere una prueba con un nivel bajo de lectura, o personas con capacidades motrices muy limitadas. Es importante ser tan específicos como sea posible acerca de estas cuestiones antes de empezar la búsqueda de la prueba.

El siguiente paso es compilar una lista inicial de candidatas. Aquí, se trata de lanzar una gran red, por lo que se pueden consultar todas las fuentes de información citadas en el capítulo 2. Es esta etapa serán de particular importancia las fuentes electrónicas, las listas exhaustivas de pruebas y otros usuarios, es decir, profesionales que usen pruebas de manera regular en el dominio en cuestión.

La tercera etapa implica restringir la lista de candidatas a un número manejable para hacer un examen más detallado. Aunque no hay un número mágico, entre 3 y 6 pruebas suelen ser suficientes. La selección de la lista final implica juzgar la calidad general de las pruebas y el grado en que cumplen con las condiciones especificadas en el paso 1.

El cuarto paso consiste en compilar información detallada sobre las pruebas finalistas del paso 3. Esto requerirá recurrir a otras fuentes, en especial, reseñas de pruebas, catálogos y personal de las editoriales y, nuevamente, otros usuarios. También es habitual obtener los materiales completos de las pruebas, por ejemplo, manuales y materiales auxiliares (p. ej., informes muestra). Ahora se dedica un tiempo considerable a examinar en detalle los materiales. Como parte de esta etapa, puede hacerse una aplicación piloto de uno o varios instrumentos; además de las puntuaciones de las pruebas, en esta aplicación se pueden observar las reacciones de los aplicadores y examinados frente a la prueba. Los resultados de esta etapa a menudo se resumen en un formato de matriz, como se ilustra en el cuadro A-4.

Cuadro A-4. Parte de una matriz en que se resumen las pruebas que se examinan en detalle durante el proceso de selección de una prueba

Característica/Prueba	**A**	**B**	**C**	**D**
Editorial				
Costo/persona				
Tipo de normas				
Confiabilidad				
Tiempo de aplicación				

Este cuadro muestra sólo algunas de las características de las pruebas que se resumen en un estudio real.

La quinta etapa consiste en hacer la selección final o, en algunas situaciones, una recomendación a algún otro individuo o grupo que se encargará de la selección. Esta decisión depende de la combinación de todos los factores relacionados con la prueba y el uso que se le quiere dar; estos factores incluyen las características psicométricas de la prueba y consideraciones prácticas como costo, ventajas y materiales de apoyo. La decisión final siempre se basa en el estándar relativo descrito antes, es decir, eliges la mejor aunque no sea perfecta.

APÉNDICE B
Cómo construir una prueba (sencilla)

Aquí presentamos un conjunto de pasos para construir una prueba relativamente sencilla (véase cuadro B-1), que quizá sigas como ejercicio en un curso. Este proyecto puede llevarse a cabo de manera individual, en pareja o en pequeños grupos. Aunque estas instrucciones no te indicarán la manera de elaborar una prueba compleja como el MMPI o el SAT, sí te ayudarán a tener práctica con los pasos importantes de la elaboración de pruebas que se bosquejaron en el capítulo 6.

Te sugerimos elegir uno de los siguientes tipos de pruebas:

1. Prueba de aprovechamiento basada en algunos capítulos de este libro.

2. Prueba de una capacidad mental específica como razonamiento cuantitativo o vocabulario.

3. Encuesta de actitud sobre un tema actual de interés.

Aunque puede ser muy tentador, no te recomendamos tratar de construir una prueba objetiva de personalidad, una prueba proyectiva, un inventario de intereses o una prueba de capacidad mental con puntuaciones múltiples.

Cuadro B-1. Resumen de los pasos para construir una prueba
1. Formular el propósito de la prueba.
2. Considerar los temas del diseño preliminar.
3. Preparar los reactivos.
4. Poner a prueba los reactivos.
5. Llevar a cabo la estandarización y los programas de investigación complementaria.
6. Preparar los materiales finales de la prueba.

Formula el propósito de tu prueba

Formula de manera concisa el propósito, incluyendo la población a la que está dirigida. En el caso de una prueba de aprovechamiento, basada en material de este libro, determina si la prueba medirá el conocimiento del material antes de que sea visto en clase, después de la lectura inicial pero antes de revisarlo en profundidad o después de un tratamiento minucioso de los capítulos. En el caso de una prueba de capacidad, sugerimos que te enfoques en un rasgo bien limitado, por ejemplo, la capacidad para completar patrones numéricos o el conocimiento de palabras. Si se trata de una encuesta de actitud, determina la amplitud con que estará definido el constructo; por ejemplo, si el tema es la pena capital, ¿la encuesta se enfocará sólo en temas legales actuales o en

temas más extensos? Si el tema es una campaña política actual, ¿la encuesta se centrará en sólo algunos candidatos o en una amplia gama de temas políticos, que quizá van de los puntos de vista conservadores a los liberales?

Nota: En la práctica real, después de formular el propósito de tu prueba, usarías las fuentes de información del capítulo 2 para determinar si ya existe una prueba que sirva a tu propósito. Sin embargo, ya que la intención de este ejercicio es construir una prueba, omitirás las fuentes de información por esta vez.

Considera los temas del diseño preliminar

Observa la lista de los temas del diseño preliminar que se discutieron en las páginas XX-XX del capítulo 6. Aquí se presentan algunas directrices para usarlas en este ejercicio; desde luego, tú puedes tomar otras decisiones. ¿Cuántas puntuaciones tendrá la prueba? Nosotros te sugerimos limitar la prueba a una sola puntuación. ¿Se aplicará de manera individual o grupal? Sugerimos la aplicación grupal. ¿Cuántos reactivos tendrá? En la versión final de la prueba, es decir, después de la selección de reactivos, sugerimos unos 30 reactivos si se trata de una prueba de aprovechamiento o de capacidad y unos 15 si se trata de una medida de actitud. ¿Qué tipo de reactivos se usarán? Sugerimos reactivos de respuesta cerrada para este ejercicio; usa reactivos de opción múltiple con tres, cuatro o cinco opciones si elaboras una prueba de aprovechamiento o de capacidad. Si es de actitudes, usa el formato tipo Likert con una escala de 3-7 puntos. Considera qué tipo de distribución de puntuaciones deseas; por ejemplo, ¿quieres una distribución con asimetría negativa con la máxima discriminación en la parte inferior? ¿Quieres una distribución aproximadamente normal de las puntuaciones? Además, considera si la aplicación de la prueba requerirá alguna adaptación especial para personas con discapacidades. Por último, realiza cualquier investigación necesaria de los antecedentes, sobre todo si se trata de una prueba de capacidad o actitud. En general, trata de que las cosas sean sencillas.

Prepara los reactivos de la prueba

Antes de empezar a escribir los reactivos, prepara un anteproyecto de la prueba en el que se bosqueje el número y tipo de reactivos que deseas. Considera el caso de una prueba de aprovechamiento basada en tres capítulos de este libro. El anteproyecto indicará el número de reactivos por capítulo y su naturaleza; por ejemplo, un tercio de los reactivos puede basarse en las palabras clave al final de cada capítulo, otro tercio puede basarse en los resúmenes de los principales puntos y el último tercio puede relacionarse con la aplicación de los conceptos. En el caso de una encuesta de actitud, el anteproyecto mostrará el rango de temas que abarcará y, tal vez, la distribución de reactivos redactados de manera positiva o negativa.

Nosotros te sugerimos preparar al menos el doble de reactivos de los que conformarán la versión final. Por ejemplo, si tu prueba de aprovechamiento será de 30 reactivos,

prepara 60; si tu encuesta de actitud será de 15 reactivos, prepara 30. Mientras haces esto, sin duda *probarás de manera informal* los reactivos para asegurarte de que sean comprensibles. Si los reactivos que preparas, por ejemplo, series de números para una prueba de razonamiento cuantitativo, son todos demasiado fáciles o demasiado difíciles para la población meta, desearás hacer ajustes antes de escribir 80 reactivos inapropiados.

Prepara las instrucciones de la prueba; deberán ser sencillas y completas. Hacer esto requerirá que decidas cómo se registrarán las respuestas, lo cual, a su vez, requerirá que pienses cómo procesarás los datos en el análisis de reactivos. Te sugerimos que los examinados registren sus respuestas en una hoja escaneable; revisa en tu centro de cómputo o de aplicación de pruebas qué tipo de hojas de respuesta escaneables están disponibles y qué tipo de resultados se pueden obtener mediante el escáner. Tu profesor puede ayudarte a contestar estas preguntas. Si no tienes acceso a un escáner, elige otro modo en que los examinados registren sus respuestas.

Mientras preparas las instrucciones, considera qué información podrías desear que proporcione el examinado además de las respuestas a los reactivos. Por ejemplo, ¿preguntarás el género, GPA actual, edad, raza, etc.? Esta información complementaria podría ser importante para describir el grupo de prueba, así como para el análisis de las puntuaciones.

Pide a un amigo que edite todos los reactivos e instrucciones para corregir cualquier error de ortografía o gramática y lograr la mayor claridad posible. Haz correcciones con base en esta revisión y, luego, prepara el borrador de la forma final para hacer una primera prueba. Ésta debe verse, en la medida de lo posible, como la forma final, con una apariencia profesional, terminada. Si la forma de prueba es demasiado extensa, considera dividirla en dos o tres formas equivalentes para poner a prueba los reactivos.

Pon a prueba los reactivos

Recuerda que poner a prueba los reactivos tiene tres fases (véase p. XX). Primero, hay una *prueba informal*. Aplica la prueba a algunos individuos representativos de la población a la que está dirigida; esto puede hacerse de manera individual o en un grupo pequeño. Pide a los examinados que comenten, de manera oral o escrita, sobre el formato, las instrucciones y los reactivos específicos. En esta prueba informal, no te interesan las puntuaciones de los examinados, sino si la tarea es clara, si el vocabulario es apropiado, etc. No debes reaccionar a los comentarios;

Si un examinado no entiende algo, no discutas con él, cámbialo. Después de hacer cambios con base en la prueba informal, estás listo para hacer la prueba formal.

Hay dos preguntas clave que necesitas responder para llevar a cabo la prueba formal. Primero, determina si necesitas la aprobación del IRB (Institutional Review Board); ya que las pruebas se aplicarán a humanos, se puede exigir esto. O tal vez tu profesor ha obtenido una aprobación en blanco para estos proyectos; o ya que se trata sólo de un

ejercicio escolar, tu institución puede no exigir ninguna aprobación. Pero si es necesaria una, consigue las formas requeridas e inicia el proceso de aprobación tan pronto como sea posible, pues la aprobación IRB puede provocar un retraso importante. Segundo, determina cuántos examinados participarán en el grupo de prueba; por lo general, es deseable tener un mínimo de 200 o, si se emplean métodos de la TRR, muchos más. Sin embargo, para un proyecto escolar, tendrás que conformarte con 20 o 30 examinados. Por último, imprime los materiales de la prueba y haz los arreglos para la aplicación real de la forma de prueba de tu prueba.

Después, obtén los *estadísticos de los reactivos*, para lo cual te recomendamos usar un programa estadístico. Otra vez, revisa en tu centro de cómputo o de aplicación de pruebas qué programa estadístico está disponible. La mayoría de las instituciones tendrá los programas que se usan en los exámenes de los cursos; dichos programas suelen producir un índice de dificultad y uno de discriminación por cada reactivo. Muchos también producen una medida de confiabilidad (casi siempre, el coeficiente alpha) y proporcionan los estadísticos descriptivos básicos de la prueba en su conjunto (media, desviación estándar, distribución de frecuencia, etc.). Si tu grupo de prueba es lo suficientemente grande, también puedes obtener los parámetros de los reactivos de la TRR, en especial el parámetro "b".

Si tu institución no cuenta con un programa estadístico diseñado específicamente para el análisis de la prueba, intenta usar un programa estándar como el SPSS. Introduce los datos: cada renglón corresponde a un examinado y cada columna a un reactivo. En el SPSS, puedes usar los siguientes comandos. En Transform, usa los comandos de Compute para generar una puntuación total de la prueba. En Analyze, usa Correlate/Bivariate, luego Options para producir Means y Standard Deviations y obtener los valores *p* y las correlaciones de cada reactivo con la puntuación total. También en Analyze, usa las rutinas Scale/Reliability para generar los coeficientes alpha y/o los de consistencia interna (división por mitades). Asegúrate de revisar las opciones en Statistics en la rutina de Reliability.

Si no puedes usar un programa de cómputo para obtener los estadísticos de los reactivos, necesitaras hacer los cálculos a mano. Si éste es el caso, te sugerimos usar una división de 50-50 de los grupos superior e inferior necesarios para obtener el índice de discriminación, suponiendo que tu grupo de prueba será relativamente pequeño (menos de 50 casos).

La fase final del análisis de reactivos es la *selección de reactivos*: aquellos que tienen las propiedades psicométricas óptimas quedarán en la prueba final. Lo que es "óptimo" depende en parte del propósito de la prueba. ¿Quieres reactivos con un nivel de dificultad concentrado alrededor de cierto punto, por ejemplo, .80 o .50? ¿Quieres niveles de dificultad distribuidos en un rango amplio?

Será necesario que *examines todos los estadísticos de los reactivos en detalle*. ¿Qué reactivo no funciona como se esperaba? ¿Algún reactivo tuvo un índice de discriminación muy bajo, quizá incluso negativo? En el caso de los reactivos de aprovechamiento, ¿alguno de los distractores tuvo un funcionamiento inusual? ¿Qué reactivos tienen los

mejores índices de discriminación? ¿Estos reactivos tienen algo en común?

Si tiene información descriptiva complementaria sobre los examinados del grupo de prueba, por ejemplo, género o edad, quizá quieras analizar los reactivos y las puntuaciones totales por cada subgrupo.

Concluye el análisis de reactivos decidiendo cuáles se mantendrán en la versión final de la prueba. Recuerda que tú elegiste la extensión de la prueba en las consideraciones del diseño preliminar. Si querías una prueba de 30 reactivos y pusiste a prueba 40, seleccionarás los 30 mejores. Quizá ahora quieres repensar tu intención original; si 35 de los 40 reactivos tienen estadísticos excelentes, puedes decidir usar los 35, con lo cual aumentarás un poco la confiabilidad de la prueba. Si sólo 25 de los 40 reactivos tienen estadísticos al menos razonables, quizá sea necesario que limites tu prueba a 25 reactivos aunque la confiabilidad sea menor de lo que habías deseado al principio.

Recuerda que en muchos proyectos de elaboración de pruebas puede haber una segunda vuelta para poner a prueba los reactivos. Es decir, si después de la primera prueba de reactivos se obtienen sólo unos cuantos de alta calidad, los autores preparan más y los ponen a prueba. Es muy poco probable que tengas tiempo para poner a prueba los reactivos por segunda vez en el contexto de un proyecto escolar.

Lleva a cabo la estandarización y los programas de investigación complementaria

Es poco probable que puedas llevar a cabo todos los pasos anteriores y realizar los programas de estandarización y de investigación complementaria en el tiempo que permite un curso académico. Quizá sea necesario que renuncies a la estandarización y continúes directamente con el sexto y último paso. Sin embargo, si el tiempo lo permite, puedes realizar los siguientes programas.

Primero, define de manera cuidadosa las características deseadas del grupo de estandarización que sea representativo de la población meta. ¿Cuáles son las características importantes de este grupo en relación con el rasgo que estás midiendo? Considera los Tipos de información para juzgar la utilidad de un grupo de normalización del cuadro 3-6. ¿Cómo puedes obtener esta información de la población meta? ¿Cómo puedes obtener esta información del grupo de estandarización? Considera cómo asegurarás la cooperación de los participantes en el proceso de estandarización.

¿Cuántas personas deben estar en la estandarización? Como señalamos en el capítulo 3, una muestra de 200 casos producirá normas muy estables. Sin embargo, es importante que la muestra sea representativa de la población meta.

¿Qué tipo de puntuaciones conformarán las normas: rangos percentiles, puntuaciones T, stanines? ¿Habrá normas por subgrupos, por ejemplo, por género o edad? La respuesta a esta última pregunta dependerá en parte del propósito de la prueba y en parte de si hay diferencias notables en el desempeño de los subgrupos en la prueba.

Prepara los cuadros que muestren las normas y los estadísticos descriptivos de

resumen. También obtén una medida de confiabilidad de consistencia interna de la muestra de estandarización. El coeficiente alpha es la elección usual.

Considera si quieres obtener la confiabilidad de test-retest. Si es así, ¿cuál será el intervalo de tiempo? ¿El estudio será realizado con la muestra entera de estandarización o con un subgrupo? ¿Quieres obtener la correlación entre las puntuaciones de tu prueba y las de otras pruebas relacionadas? Si es así, ¿cómo obtendrás esta información?

Prepara los materiales de la prueba final, incluyendo el manual

Sin importar si se llevó a cabo la estandarización, debes preparar los materiales y el manual de la prueba final. Los materiales suelen consistir en un cuadernillo de la prueba y algún tipo de documento para las respuestas. El cuadernillo de la prueba debe tener una apariencia profesional, es decir, sobria, clara y sencilla, sin brillo ni gráficos disparatados.

El manual debe identificar el propósito de la prueba y describir los fundamentos de su elaboración, incluyendo la discusión de los temas del diseño preliminar. También debe ofrecer detalles relacionados con la elaboración y prueba de reactivos, y describir la muestra con la que éstos se pusieron a prueba. El manual debe presentar resúmenes de los estadísticos de los reactivos y de la prueba total. Si no se realizó la estandarización, los estadísticos serán los que se obtuvieron en el programa de análisis de reactivos. Si se realizó la estandarización, los estadísticos serán los que se obtuvieron en el programa de estandarización, incluyendo la descripción de la muestra. Debe resumirse la información sobre la confiabilidad, incluyendo el error estándar de medición. El manual también debe resumir los datos de cualquier análisis de subgrupos que se haya hecho, por ejemplo, diferencias por género. Si realizas todos estos pasos, o incluso la mayoría, sin duda aprenderás mucho acerca de la empresa de las pruebas psicológicas.

Las 10 cosas que he aprendido sobre la elaboración de pruebas

He trabajado minuciosamente en proyectos de elaboración de pruebas durante casi toda mi vida profesional. Este trabajo ha incluido proyectos grandes (de millones de dólares y de varios años) y pequeños, comerciales y no comerciales, de pruebas cognitivas y no cognitivas. Ha incluido cada fase de la elaboración de pruebas, desde la conceptualización original hasta la publicación final y la investigación posterior. De todas estas experiencias, pienso que he aprendido algunas cosas que van más allá del tratamiento de este tema en un libro de texto estándar. Puede ser que los estudiantes no puedan aprender estas cosas sin experimentar por sí mismos el trabajo de elaboración de pruebas, pero sentiría que es un acto de negligencia no consignar estos pensamientos con la esperanza de que sean útiles a los estudiantes. Así que aquí está mi lista con las 10

cosas que creo haber aprendido sobre la elaboración de pruebas.

1. La conceptualización original es más importante que el trabajo técnico/estadístico. En los libros de texto, los tratamientos de la elaboración de pruebas tienden a concentrarse en los procesos de escribir reactivos y los procedimientos de análisis de reactivos. Algunos libros de texto ni siquiera mencionan las etapas previas de definir el propósito de la prueba y las consideraciones del diseño preliminar. En mi experiencia, la conceptualización original de la prueba es mucho más importante que el trabajo técnico/estadístico. Si no tienes una conceptualización clara, ni la más fina redacción de los reactivos, ni un gran dominio de la estadística te salvarán en esta empresa.

2. Necesitas dedicar mucho tiempo al estudio del área antes de empezar a escribir los reactivos. Conocer todos los pasos de la construcción de pruebas –incluso conocerlos muy bien– no te califica para empezar a escribir una prueba. Es esencial que dediques cierto tiempo a estudiar el área en la que quieres hacer tu prueba. Si se trata de habilidades de lectura, necesitas saber qué dicen los estudios de investigación y los expertos acerca de las habilidades de lectura. Si deseas medir ansiedad, necesitas conocer la literatura de esa área.

3. En la etapa del diseño original, necesitas pensar en los informes finales de las puntuaciones. Lo que en realidad estás produciendo no es una prueba, sino una puntuación que se asigna a alguna persona. ¿Qué es exactamente lo que le ofrecerás a esa persona? ¿Cómo se verá la puntuación final? Si esperas hasta estar listo para publicar la prueba antes de pensar sobre los informes de las puntuaciones, casi sin duda encontrarás que habrías deseado haber construido la prueba de otra manera.

4. Al preparar los reactivos, busca la sencillez. Los reactivos "ingeniosos" a menudo no funcionan. Los reactivos sencillos casi siempre funcionan mejor que los complejos. Si escribes un reactivo que piensas que en verdad es ingenioso, es muy probable que sea confuso para los examinados y, por lo tanto, sus estadísticos serán pobres.

5. Asegúrate de poner a prueba suficientes reactivos: por lo general, el doble de los necesarios para la versión final. Es un fastidio escribir muchos reactivos; en realidad, es un trabajo aburrido. Además, al hacerlo, casi estás seguro de que es un buen reactivo, así que hay una gran tentación de preparar sólo unos cuantos más de los que quieres en la versión final. El hecho es que incluso con una buena prueba informal y una edición amplia de los reactivos perderás muchos de ellos en la etapa del análisis formal de los reactivos. Asegúrate de incluir suficientes en tu programa de análisis de reactivos.

6. Haz una prueba informal sencilla antes de la prueba principal. Las pruebas informales, sencillas son muy fáciles de hacer. Puedes recurrir a amigos, colegas o cualquier persona que esté disponible. Es sorprendente la frecuencia con que un reactivo que no ha pasado por una prueba informal se incluye en un gran estudio nacional sólo para esfumarse como el viento cuando incluso en la prueba informal más sencilla habría podido identificarse un defecto fundamental en él.

7. Los "malos" reactivos casi siempre son fáciles de reconocer. Los estadísticos del análisis de reactivos no hacen mucha diferencia. Existen varios estadísticos del análisis de reactivos. Podemos leer interminablemente acerca de las ventajas de éste o aquel método en relación con los demás. Desde luego, tú quieres usar el mejor o más apropiado; sin embargo, en mi experiencia, la metodología específica del análisis de reactivos no hace mucha diferencia. Los "malos" reactivos suelen ser fáciles de reconocer con cualquier método. Como corolario de este punto, debo señalar que jugar con distractores basados en los datos del análisis de reactivos, por lo general, no determina a un mal reactivo. En general, es preferible escribir uno nuevo o usar uno diferente.

8. Desde el punto de vista estadístico, el grupo de estandarización no tiene que ser grande, si se selecciona de manera adecuada. No obstante, las personas se impresionan principalmente con el tamaño del grupo. Respecto de un grupo de estandarización, la primera pregunta de los usuarios, sin excepción, es acerca del tamaño del grupo. Ésta es una pregunta equivocada. Lo importante es si el grupo es representativo de la población meta o, al menos, de alguna población bien definida. Podemos obtener normas muy estables con sólo unos cientos de casos.

9. Por favor, termina el manual definitivo. Por razones que no son del todo claras, parece ser muy difícil poner punto final al manual de la prueba. A veces se publica una prueba con un manual, instrucciones o diversos informes técnicos provisionales.

Tal vez, los autores de pruebas están exhaustos para cuando terminan los programas de estandarización y de investigación complementaria. Cualquiera que sea la razón, por favor, termina el manual definitivo de la prueba.

10. El proceso completo siempre toma más tiempo del que pensaste. Al inicio del proyecto de elaboración de pruebas, éste parece muy factible. El entusiasmo está por los cielos. La anticipación de tener una prueba nueva y buena es estimulante. Se escribirá con rapidez los reactivos, las personas participarán con gusto en los programas de análisis de reactivos y de estandarización, y el manual definitivo ya está bosquejado adecuadamente en tu mente, o así parece. Pero, ¡ay!, la realidad es muy diferente. Un consejo: haz un cronograma para tu proyecto, pero da por hecho que, por desgracia, no cumplirás al menos con algunos plazos de tu cronograma.

APÉNDICE C
Información de contacto de las principales editoriales de pruebas

Cientos de compañías e incluso individuos publican pruebas. Diversas fuentes, como las citadas en el capítulo 2, mantienen directorios exhaustivos de estas editoriales, de los cuales el del Buros Institute es el más destacado, pues incluye más de 900 entradas. Este directorio está disponible en versión impresa en las ediciones actuales de *Tests in Print* y el *Mental Measurements Yearbook*.

A continuación presentamos una lista con la información de contacto sólo de las editoriales citadas en este texto. Aunque es una lista relativamente pequeña, estas editoriales producen quizá 95% de todas las pruebas que se usan en EUA y cubren casi todas las pruebas que mencionamos en este libro.

La información más importante para los estudiantes, además del nombre de las editoriales, es la dirección de su página de internet. Con ella, pueden obtener información muy útil acerca de las pruebas citadas en el texto. Anteriormente, era posible terminar un curso de pruebas psicológicas sin ver nunca una prueba o un informe de las puntuaciones. En cambio, hoy la experiencia de los estudiantes se ha enriquecido considerablemente por tener acceso a información de las pruebas mediante las páginas de internet de las editoriales. Los alentamos a acceder a esta información con regularidad. Desde luego, deben darse cuenta de que las editoriales están en el negocio de comercializar sus productos, por lo que las afirmaciones acerca de sus pruebas deben evaluarse con cuidado. Las URL enumeradas aquí pueden cambiar, pero estas editoriales se pueden rastrear con facilidad con ayuda de cualquier buscador. Como otras partes del mundo de los negocios editoriales, con frecuencia las editoriales de pruebas son vendidas y compradas por una u otra empresa. Así, lo que en 2013 fue la editorial ABC, ahora puede ser parte de la editorial XYZ. Los buscadores hacen un trabajo sorprendente al ponerte en contacto con la fuente correcta.

Editorial	Página de internet
ACT, Inc.	www.act.org
AGS Publishing (American Guidance Service)	www.ags.net
College Board	www.collegeboard.com
CPP (Consulting Psychologists Press)	www.cpp-db.com
CTB/McGraw-Hill	www.ctb.com
EDITS	www.edits.net

Educational Testing Service (ETS)	www.ets.org
Harcourt Assessment	www.harcourtassessment.com
Institute for Personality and Ability Testing (IPAT)	www.ipat.com
Kuder, Inc.	www.kuder.com
Lafayette Instruments	www.lafayetteinstrument.com
Mind Garden	www.mindgarden.com
National Occupational Competency Testing Institute (NOCTI)	www.nocti.org
Pearson	www.pearsonassessments.com
PRO-ED	www.proedinc.com
PAR (Psychological Assessment Resources)	www.parinc.com
Psychological Corporation	www.harcourtassessment.com
Riverside Publishing	www.riverpub.com
Western Psychological Services (WPS)	www.wpspublish.com

APÉNDICE D
Conjuntos de datos muestra

El apéndice D contiene conjuntos de datos descargables de la página de internet de la editorial que corresponden a este libro. Por favor, consulta en www.wiley.com las secciones Companion para profesores y alumnos (Instructors, Students) que corresponden a *Psychological Testing*, Third Edition, de Hogan, donde encontrarás conjuntos de datos. El primero (D1) contiene los GPA y diversas posibles variables predictoras, incluyendo las puntuaciones del SAT y el NEO PI. El segundo (D2) contiene datos de un estudio sencillo de confiabilidad. El tercero (D3) contiene datos de reactivos adecuados para llevar a cabo los procedimientos del análisis de reactivos. Los datos están disponibles para IBM, SPSS y Microsoft Excel. Los archivos incluyen una descripción detallada de cada conjunto de datos. Por último, hay una hoja de Excel llamada ICC Generator acompañada por un documento de Word etiquetado como "Instructions for Using the ICC Generator". Abre estos documentos y sigue las instrucciones para generar las curvas características del reactivo (CCR) variando los tres parámetros de un modelo de la TRR.

APÉNDICE E
Respuestas de ejercicios seleccionados

A continuación aparecen las respuestas de ejercicios seleccionados que se encuentran en el texto de los capítulos o en la parte final de cada uno. Muchos ejercicios conceden una considerable libertad a las respuestas de los estudiantes, pero en este apéndice no se tratan esos ejercicios. En algunos casos se incluyen notas sobre cómo responder un ejercicio.

Capítulo 1

Inténtalo en la página 7: EDI = Eating Disorder Inventory, GRE = Graduate Record Examination, WAIS = Wechsler Adult Intelligence Scale (Escala Wechsler de Inteligencia para Adultos), MMPI = Minnesota Multiphasic Personality Inventory (Inventario Multifásico de Personalidad de Minnesota).

Ejercicio 5: Binet dijo que él intentaba obtener una medida de inteligencia general, sin importar las causas precisas del funcionamiento mental actual, y distinguió esta capacidad de los problemas conductuales.

Ejercicio 10: Piaget fue alumno de Simon (colaborador de Binet). La obra de Piaget influyó en los teóricos contemporáneos Sternberg y Gardner. Nosotros retomamos estos autores en el capítulo 7.

Capítulo 2

Todos los *Ejercicios* de este capítulo requieren usar páginas de internet y los recursos de la biblioteca, por lo cual las respuestas diferirán dependiendo de las URL elegidas y tu biblioteca local.

Capítulo 3

Inténtalo de la página 68: Media = 4.0, mediana = 4, moda = 5.

Inténtalo de la página 84: CI de razón de Matt = 93, de Meg = 155.

Ejercicio 2: $M = 6.00$, Mediana = 6, $DE = 2.12$.

Nota: Los estudiantes pueden obtener respuestas ligeramente diferentes de la *DE*

según el programa de cómputo que usen. Si se usa N en vez de $N - 1$ en el denominador para obtener la *DE*, entonces *DE* = 1.90.

Ejercicio 5:

$z = +1.0$	Percentil = 84	NCE = 71	CI Wechsler = 115
Percentil = 75	$z = .67$ (aproximadamente)	Otis-Lennon = 111	Stanine = 6
Puntuación T = 30	Percentil = 2	Stanine = 1	Puntuación $z = -2.0$

Ejercicio 8: Para una puntuación natural = 59, percentil = 86 y CID = 117.

Capítulo 4

Inténtalo de la página 113: GPA predicho = 3.29.

Inténtalo de la página 136: Con $r = .92$, EEM = 4.2. Con $r = .70$, EEM = 8.2.

Ejercicio 1: $r = .96$.

Ejercicio 4: Con 20 reactivos y una confiabilidad original (r_o) = .75, obtén la confiabilidad corregida (r_c) mediante la fórmula 4-13. Cuadruplicando la extensión de la prueba ($n = 4$), se obtiene $r_c = .92$. Cuando se reduce a la mitad la extensión de la prueba ($n = 1/2$), se obtiene $r_c = .60$. Para obtener $r_c = .90$, se requiere $n = 3$. Por lo tanto, la extensión triplicada equivale a $20 \times 3 = 60$ reactivos.

Ejercicio 10: Para *DE* = 10, $r = .80$, $EE_{dif} = 6.32$.

Para *DE* = 10, $r = .60$, $EE_{dif} = 8.94$.

Capítulo 5

Inténtalo de la página 153: El ejemplo B muestra la mejor validez. Los ejemplos A y C muestran la mayor cantidad de varianza irrelevante para el constructo.

Inténtalo de la página 164: $EE_{est} = 1.89$.

Ejercicio 3: GPA = 3.60, $EE_{est} = .32$, probabilidad de 11%.

Ejercicio 6: Usa la fórmula 5-4. La *r* corregida de la prueba A con el GPA es .58; de la prueba B con el GPA es .51.

Capítulo 6

Inténtalo de la página 219: Prop. correcta = .85, Índice de disc. = .23.

Inténtalo de la página 223: La CCR del reactivo A tiene una pendiente pronunciada y cruza la marca de 50% de probabilidad en el eje de Y arriba de −1 en el eje X. La CCR del reactivo B se eleva ligeramente y cruza la marca de 50% de probabilidad en el eje Y arriba de 2 en el eje X.

Inténtalo de la página 227: .20 0 .00 40 .40 .40

Ejercicio 5: En el cuadro 6-9, el valor p del reactivo 10 es .62. El porcentaje (o Prop.) correcto para el reactivo 23 en el grupo inferior es .00. El reactivo más fácil es el 28. La diferencia entre porcentaje correcto en los grupos superior e inferior para el reactivo 29 es .05.

Ejercicio 8: Se parece más a la prueba A.

Capítulo 7

Ejercicio 5: El pensamiento divergente corresponde a la capacidad amplia de recuperación.

Ejercicio 8: El tamaño del efecto estimado (*d*) es .3 aproximadamente, es decir, cerca de 3/10 de una desviación estándar.

Capítulo 8

Inténtalo de la página 297:

46 13 13

33 7 13

Ejercicio 9: Las distribuciones de Vocabulario se superpondrán casi por completo con el grupo de edad 20-34 desplazado ligeramente a la derecha. En el caso de Diseño con cubos, la media del grupo de edad 85-89 estará centrada alrededor del percentil 5 del grupo de edad 20-34.

Capítulo 9

Inténtalo de la página 341: El ICE Total del alumno es 104. Ya que, en el OLSAT, *M* = 100 y *DE* = 16, dicho ICE corresponde a una puntuación z de .25. Para hacer otras conversiones, consulta el cuadro 3-1.

Ejercicio 5: ICE Total = 104, Verbal = 108, No Verbal = 101. Observa que las bandas de confianza de las puntuaciones, representadas de manera gráfica a la derecha del cuadro, se superponen casi por completo. Así, no intentaríamos que el alumno dé mucha importancia a la diferencia entre las puntuaciones Verbal y No Verbal.

Ejercicio 8: Si empiezas con una prueba de 15 reactivos que tiene una confiabilidad de .50, necesitas aumentar la extensión de la prueba multiplicando 15 por un factor de 5.8 (es decir, aumentar a 87 reactivos) para obtener una confiabilidad de .85. Se obtiene 5.8 como valor de *n* mediante la fórmula 4-13.

Capítulo 10

Ejercicio 4: Promedio = 10. F en Vocabulario e Información. D en Retención de dígitos, Aritmética y Claves.

Capítulo 11

Ejercicio 3: Pruebas de matemáticas del nivel P3: Solución de problemas, Procedimientos, Intermedio total 2; la Batería Completa tiene 372 reactivos. La aplicación de la prueba Avanzado 1 (A1) Ciencia es de 25 min; para el inicio de 2 grado se recomienda Primario 1 (P1).

Capítulo 12

Inténtalo de la página 443: Escala de Autoconcepto de Tennessee, cuadrante superior derecho; Inventario Básico de Personalidad, cuadrante superior izquierdo; Prueba de Chequeo de Abuso de Drogas, cuadrante inferior derecho.

Inténtalo de la página 449: Las respuestas de la persona C son inconsistentes. No puede ocurrir que de verdad estés en desacuerdo con ambas afirmaciones.

Ejercicio 6: Falseamiento positivo: V, V, F, F.

Capítulo 13

Inténtalo de la página 498: Las escalas Impulso por la delgadez y Baja autoestima muestran la mayor separación. Las escalas Perfeccionismo y Miedo a la madurez muestran la menor separación.

Ejercicio 1: Código de dos puntos: 42 (es decir, las dos puntuaciones T más altas).

Capítulo 14

Ejercicio 6: Con $M = 130$, $DE = 17$, una puntuación de 145 corresponde a una puntuación $z = .88$. Usando un cuadro de curva normal, encontramos que aproximadamente 19% de los casos está arriba de $z = .88$.

Capítulo 15

Inténtalo de la página 563: La puntuación de Dick es 0. La de Tom es +20. La de Harry es −20.

Ejercicio 2: Las respuestas muestran bastante bien, aunque lejos de ser perfecta, una

concordancia interna. Tom está de acuerdo o totalmente de acuerdo con la mayoría de los reactivos. Harry está en desacuerdo o en total desacuerdo con la mayoría de ellos. Dick tiende a dar respuestas de nivel intermedio.

Capítulo 16

Inténtalo de la página 580: Stanford, A; WAIS, C; Rotter, C; Strong, B.

Ejercicio 5: Al volver a calcular TS/Tasa más alta del cuadro 16-9 con los cambios indicados en la contratación, encontramos que las nuevas tasas son blancos = .80, negros = 1.00, asiáticos = .62 e hispanos = .54. Por lo tanto, ahora hay impacto adverso para los asiáticos y los hispanos.

GLOSARIO

AAIDD American Association on Intellectual and Developmental Disabilities, el principal grupo profesional que se ocupa de la definición de discapacidad intelectual y de los servicios para las personas que la padecen (retraso mental). Esta organización se llamó antes American Association on Mental Retardation (AAMR).

Acierto Caso que rebasa los puntos de corte del criterio y la prueba, o que se queda por debajo en ambos.

ACT American College Testing. Se refiere a la evaluación ACT (de admisión a la universidad), así como a la compañía que produce esa prueba, ACT, Inc.

ADA Americans with Disabilities Act [Ley de Estadounidenses con Discapacidades] de 1990; ley federal que define las discapacidades y estipula los requerimientos para las adaptaciones.

Adaptación Realizar un cambio en el ambiente o en las condiciones de la aplicación diseñado para eliminar los efectos en el desempeño en la prueba de una discapacidad y poder medir, a pesar de ello, el mismo constructo.

Afasia Déficit en la capacidad para expresar o comprender la comunicación escrita o hablada como resultado de una lesión cerebral.

AFQT *Armed Forces Qualifying Test*, prueba que se usó en algún tiempo para evaluar a los reclutas militares, que ahora ha sido reemplazada por el ASVAB. Las iniciales también hacen referencia a una puntuación compuesta derivada del ASVAB.

Alpha de Cronbach Véase **Coeficiente alpha**.

Análisis de escalograma Nombre técnico del escalamiento Guttman.

Análisis de puesto Análisis detallado de los requerimientos de un puesto, sobre todo para construir o demostrar la validez de una prueba usada para seleccionar empleados.

Análisis de reactivos Análisis estadístico de los reactivos individuales de una prueba, en especial para determinar su nivel de dificultad y poder de discriminación.

Análisis factorial Clase de método estadístico para identificar las dimensiones que subyacen en muchas puntuaciones o en otros indicadores del desempeño.

Análisis multirrasgo-multimétodo Técnica para examinar las relaciones entre distintas variables, cada una medida de distintas maneras.

Anteproyecto de una prueba Bosquejo del contenido de una prueba, en especial aplicado a las pruebas de aprovechamiento cuando se elaboran mediante el análisis de materiales curriculares y requerimientos del puesto.

Apraxia construccional Incapacidad para armar o copiar objetos bi o tridimensionales.

Asimetría Asimetría en una distribución; en la asimetría a la izquierda o asimetría

negativa, las puntuaciones se amontonan en la parte alta de la distribución y resulta una larga cola a la izquierda; en la asimetría a la derecha o asimetría positiva, las puntuaciones se amontonan en la parte baja de la distribución y resulta una larga cola a la derecha.

ASVAB *Armed Services Vocational Aptitude Battery* [Batería de Aptitudes Vocacionales de los Servicios Armados].

Atenuación Disminución o reducción; en el campo de las pruebas se refiere a la reducción de la correlación entre dos variables debido a la imperfecta confiabilidad y/o a la homogeneidad del grupo.

Atkins Proceso judicial de *Atkins contra Virginia* que llevó al fallo de que las personas con retraso mental no estaban sujetas a la pena capital.

Automonitoreo Registros cuidadosos y detallados que la propia persona lleva de sus conductas y las condiciones que las rodean.

Banda de confianza Banda ubicada alrededor de una puntuación que se basa en el error estándar de medición.

Batería Conjunto de pruebas coordinadas que cubren diferentes áreas de contenido y niveles de edad/grado. Por lo común, esta palabra se aplica a pruebas estandarizadas de aprovechamiento.

Batería fija Conjunto de pruebas neuropsicológicas que se aplican, en su totalidad, a cada examinado.

Batería flexible Conjunto de pruebas neuropsicológicas del que se aplican algunas dependiendo del caso, por lo que no se aplican las mismas pruebas a cada examinado.

Batería psicoeducativa Conjunto de pruebas aplicadas individualmente para evaluar las capacidades mentales y el aprovechamiento de manera coordinada, sobre todo en relación con las dificultades de aprendizaje, TDAH, etc.

BDI Beck Depression Inventory [Inventario de Depresión de Beck], la medida de depresión más usada.

Binet, Alfred Psicólogo francés que creó la primera medida viable de capacidad mental general, la escala Binet-Simon, que condujo a la Stanford-Binet; el término Binet también se usa para designar las pruebas mismas.

Broca Cirujano francés que fue el primero en documentar el sitio del daño cerebral asociado con una incapacidad para hablar, pero con la comprensión del lenguaje intacta; también designa el área afectada del cerebro.

Buros Puede referirse a la persona (Oscar Krisen Buros), el instituto (Buros Center for Testing) o, con mayor frecuencia, a la serie de reseñas de pruebas publicadas (Buros *Mental Measurements Yearbook* [véase **MMY**]).

Calificación analítica Calificación de un ejercicio de una prueba de varios rasgos o características que suponemos diferentes.

Calificación automatizada Calificación del desempeño en ejercicios complejos mediante un análisis por computadora.

Calificación holística Asignar una sola puntuación a un ensayo (o una tarea similar) con base en la impresión general de su calidad; opuesta a la calificación analítica.

Cambio real Cambio real en un rasgo o característica subyacente, en contraste con las fluctuaciones momentáneas.

Capacidades mentales primarias Teoría de la inteligencia de factores múltiples de Thurstone que sugiere que hay cerca de siete dimensiones distintas de la capacidad mental.

Cattell James McKeen Cattell, pionero estadounidense del desarrollo de la teoría y los métodos de las pruebas; a menudo se le reconoce como el padre del campo de las pruebas mentales.

Cero verdadero Punto de una escala que representa la ausencia total de cantidad; contrasta con un cero arbitrario como, por ejemplo, en la escala Farenheit.

Certificación Procedimiento para demostrar que una persona cumple con los requisitos para algún puesto u otro tipo de posición; indicador de que la demostración se ha llevado a cabo.

CFR Code of Federal Regulations [Código de Regulaciones Federales]; lista exhaustiva de las regulaciones que emanan del gobierno federal.

CI de desviación Normas de los CI basadas en puntuaciones estándar, por lo general con $M = 100$ y $DE = 15$.

CI de razón (EM/EC) × 100. Edad mental dividida entre la edad cronológica cuyo resultado se multiplica por 100. Es la manera anticuada de determinar el CI.

CIE CI de Ejecución de una de las escalas Wechsler de inteligencia.

CIE Clasificación Internacional de Enfermedades, elaborada por la Organización Mundial de la Salud, que ofrece una clasificación oficial de los trastornos mentales, alternativa a la del DSM.

CIT CI Total; puntuación total que combina subpuntuaciones de cada una de las escalas Wechsler de inteligencia.

CIV CI Verbal en una de las escalas Wechsler de inteligencia.

Codificación Sistema o método para asignar categorías o puntuaciones numéricas a las respuestas en una prueba proyectiva.

Código ético de la APA Nombre corto de *Ethical Principles of Psychologists and Code of Conduct* [Principios Éticos del Psicólogo y Código de Conducta] de la American Psychological Association.

Códigos de ubicación Códigos (puntuaciones) del Rorschach que indican a qué partes de las manchas de tinta una persona hace referencia en sus respuestas.

Coeficiente alpha Medida de consistencia interna de los reactivos de una prueba; a menudo se denomina **alpha de Cronbach**.

Coeficiente de correlación Expresión numérica, que va de −1.00 a +1.00, de la relación entre dos variables.

Coeficiente de correlación intraclase Tipo de correlación que expresa el grado de acuerdo entre más de dos jueces.

Coeficiente de validez Validez indicada por un coeficiente de correlación; puntuaciones de la prueba que se correlacionan con otro criterio.

College Board Antes College Entrance Examination Board (CEEB), patrocinador del SAT, así como de otras pruebas y servicios para estudiantes que irán a la universidad.

Competencia Desarrollar y mantener las habilidades y conocimientos profesionales; ejercer sólo en áreas y con técnicas en las que uno tiene tales habilidades y conocimientos.

Competencia para comparecer ante un juicio Capacidad mental apropiada en el momento de comparecer ante un juicio; contrasta con la demencia.

Conducta adaptada Conductas relacionadas con el afrontamiento de la vida cotidiana.

Confiabilidad Consistencia o fiabilidad del desempeño en la prueba en diferentes ocasiones, calificadores y contenido específico.

Confiabilidad de división por mitades Medida de confiabilidad basada en la división de la prueba en dos mitades y, luego, en la correlación del desempeño entre ellas.

Confiabilidad de formas alternas Confiabilidad determinada por la correlación de dos formas de una prueba.

Confiabilidad de pares y nones Método para determinar la confiabilidad calificando por separado los reactivos pares y los nones de una prueba.

Confiabilidad de test-retest Confiabilidad determinada al correlacionar el desempeño en una prueba aplicada en dos ocasiones diferentes.

Confiabilidad interjueces Grado de acuerdo acerca del desempeño de los individuos entre distintas personas que califican una prueba.

Confidencialidad Usar la información obtenida en un contexto profesional sólo para propósitos profesionales y con el consentimiento del cliente.

Conocimiento de los resultados Ofrecer a un cliente la revelación completa de los resultados de la evaluación.

Consentimiento informado Acuerdo de una persona para participar en la aplicación de una prueba, un tratamiento o un proyecto de investigación basado en la comprensión razonable de los riesgos y beneficios.

Consistencia de las respuestas Contestar reactivos, en especial los de un inventario de personalidad, de una manera razonablemente consistente.

Consistencia interna Reactivos que, en su mayor parte, miden el mismo rasgo o característica como lo indican las intercorrelaciones entre los reactivos. Varios métodos muestran la consistencia interna como forma de la confiabilidad de la prueba.

Constructo Rasgo o variable psicológica.

Contaminación del criterio Ocurre cuando las puntuaciones de la prueba influyen de manera constante en el criterio en un estudio de validez de criterio.

Continuo capacidad–aprovechamiento Continuo teórico que va de la capacidad o aptitud puras a habilidades aprendidas muy específicas y conocimiento; se usa para conceptualizar la diferencia entre pruebas de capacidad y de aprovechamiento.

Corrección Spearman-Brown Fórmula que permite estimar el efecto del aumento o disminución del número de reactivos en la confiabilidad.

Correlación múltiple Técnicas para combinar información de distintas variables de manera óptima para producir la mejor predicción de alguna otra variable; correlación que resulta de la información combinada y la otra variable.

Crawford Demond Crawford, demandante principal en el caso federal *Crawford contra Honig*, que estuvo relacionado con el uso de pruebas de CI en las escuelas de California.

Criterio externo Criterio, por ejemplo, desempeño escolar o laboral, usado para demostrar la validez de una prueba.

Criterio meta Seleccionar reactivos con base únicamente en si discriminan un grupo de otro.

Criterio Variable externa con la que se compara el desempeño en una prueba.

Curtosis Grado en que una "curva de campana" es puntiaguda o plana.

Curva característica del reactivo (CCR) Función matemática que describe la relación entre la probabilidad de una respuesta correcta en un reactivo y el rasgo teórico subyacente.

Curva normal Curva de densidad simétrica y unimodal con colas asintóticas; a menudo se denomina curva de campana.

DAP Draw-a-Person [Dibuja una persona]; designa una técnica general y el nombre de diversas pruebas específicas en las que el examinado dibuja una o más personas.

Debra P. Demandante principal en el caso federal *Debra P. contra Turlington* sobre el uso de una prueba para la graduación de una escuela de bachillerato en Florida.

Decimocuarta Enmienda Enmienda a la Constitución de EUA, adoptada en 1868, que incorpora las disposiciones del debido proceso y la protección igualitaria.

Demencia Capacidad mental o trastorno al momento de cometer un crimen; se contrasta con la competencia para comparecer ante un juicio.

Desatención espacial Deterioro neurológico en el que una persona no informa ver objetos en cierta parte de su campo visual, por ejemplo, en el del ojo izquierdo.

Desempeño máximo Lo mejor que una persona puede hacer cuando no hay restricciones de tiempo rígidas; a menudo se usa el término en relación con las pruebas de poder.

Desempeño típico Desempeño típico o normal para un individuo; suele contrastarse con el desempeño máximo, es decir, lo mejor que una persona puede hacer.

Desviación estándar Raíz cuadrada de la suma de las desviaciones respecto de la media (al cuadrado), dividida entre *N*; es la medida de variabilidad más común.

Determinantes Factores que presuntamente conducen a ciertos tipos de respuestas, sobre todo en el Rorschach.

DFH Dibujos de la figura humana; otro término para las técnicas draw-a-person [dibuja una persona] o draw-a-man [dibuja un hombre].

Diferencial semántico Método para valorar un objeto en una serie de escalas bipolares.

Dificultad del reactivo Porcentaje de respuestas correctas (o en una dirección específica) en un reactivo de una prueba.

Dinamómetro Artefacto para medir la fuerza de agarre.

Dirección de las respuestas Tendencia a contestar los reactivos de una prueba de personalidad en cierta dirección o con cierta disposición, sobre todo no relacionada con el rasgo que la prueba pretende medir.

Direccionalidad Dirección o tono positivo o negativo en el tronco de los reactivos de actitud.

Directrices EEOC Directrices de la Comisión para la Igualdad de Oportunidades en el Empleo [Equal Employment Opportunity Commission Guidelines] sobre el uso de pruebas cuyo fin es contratar empleados.

Discalculia Incapacidad para trabajar con cantidades numéricas, como en los cálculos sencillos.

Discapacidad intelectual Padecimiento caracterizado por una capacidad mental y una conducta adaptativa considerablemente por debajo del promedio, cuyo inicio ocurrió durante los años de desarrollo. Antes se conocía como retraso mental.

Discriminación del reactivo Índice del grado de separación entre los grupos superior e inferior en un reactivo de una prueba.

Discusión de grupo sin líder Método para observar cómo reacciona una persona en una discusión grupal relativamente no estructurada.

Dislexia diseidética Incapacidad para leer las palabras como un todo, de modo que la persona debe pronunciar todos los sonidos de la palabra.

Dislexia disfonética Incapacidad para pronunciar todos los sonidos, de modo que la persona lee palabras completas y depende de su vocabulario visual.

Dispersograma Véase **Distribución bivariada**.

Disposición de protección igualitaria Disposición de la Decimocuarta Enmienda a la Constitución de EUA que garantiza a todos los ciudadanos el derecho a la protección igualitaria al amparo de la ley.

Disposición del debido proceso Disposición de la Decimocuarta Enmienda a la Constitución de EUA que garantiza a todos los ciudadanos el derecho al debido proceso ante la ley.

Distorsión de la respuesta Tendencia a responder los reactivos de las pruebas de personalidad sin expresar los verdaderos sentimientos.

Distractor Opción en un reactivo de opción múltiple diferente a la respuesta correcta o

meta.

Distribución bivariada Representación de la relación entre dos variables en un sistema de coordenadas cartesianas; también se denomina dispersograma.

Distribución de frecuencias Distribución de puntuaciones naturales, por lo común presentada en intervalos agrupados y ordenados de los altos a los bajos.

Distribuciones superpuestas Presentación de distribuciones de frecuencia de puntuaciones de dos o más grupos que muestran un grado de superposición.

DSM *Manual Diagnóstico y Estadístico de los Trastornos Mentales* (Diagnostic and Statistical Manual of Mental Disorders), la principal publicación de la American Psychiatric Association, en la que se definen los trastornos psicológicos.

Ecuación de regresión múltiple Ecuación de regresión (predicción) que resulta de un estudio de correlación múltiple.

Edad cronológica (EC) Edad de una persona, por lo general presentada en años y meses (p. ej., 8–4 significa 8 años y 4 meses).

Edad mental (EM) Puntuación típica en una prueba de personas de determinada edad; tipo de norma de pruebas que emplea estas puntuaciones típicas.

EDI *Eating Disorder Inventory* [Inventario de Trastornos de la Conducta Alimentaria], medida ampliamente usada de los trastornos de la conducta alimentaria.

Educación basada en estándares Enfoque de la educación que surge del movimiento de responsabilidad, el cual hace hincapié en la clara identificación del contenido que se debe aprender, la especificación de los niveles de desempeño requeridos y la certeza de que todos tengan la oportunidad de aprender.

Efecto Barnum Llamado así por el famoso promotor de circo P. T. Barnum, este término se refiere a las descripciones rimbombantes que son ciertas para casi cualquier persona, pero se presentan de modo que parecen ser específicas de una persona en particular.

Efecto de Stroop Nombrar con mayor lentitud los colores de la tinta cuando está escrita una palabra incongruente con el nombre del color.

Efecto Flynn Aumento notable en el promedio de las puntuaciones de las pruebas en poblaciones enteras a lo largo de sucesivas generaciones, nombrado así por James Flynn, quien promulgó los resultados.

Encogimiento de la validez Reducción en la validez que resulta de la validación cruzada en un nuevo grupo.

Entrevista clínica estructurada Entrevista clínica que hace hincapié en el uso de las mismas preguntas y métodos de calificación con todos los clientes. Contrasta con la entrevista no estructurada tradicional, en la que los temas, preguntas y valoraciones varían con cada cliente y clínico.

EPPS *Edwards Personal Preference Schedule* [Inventario de Preferencias Personales de Edwards].

Equivalente de edad Tipo de norma en que la puntuación de una persona se compara

con puntuaciones típicas de otras personas en varios niveles de edad.

Equivalente de grado Tipo de norma de las pruebas que expresa el desempeño de una persona en comparación con el de estudiantes de distintos grados escolares.

Equivalente de la curva normal (ECN) Tipo de norma de pruebas equivalente a las normas percentiles en los percentiles 1, 50 y 99, pero tiene intervalos iguales a lo largo de toda la escala.

Error constante Error que hace que las puntuaciones sean de manera consistente más altas o más bajas de lo que merece un individuo o un grupo debido a factores no relacionados con el propósito de la prueba.

Error estándar de estimación Índice del grado de error en la predicción de una variable con base en otra.

Error estándar de la diferencia Índice de la variabilidad de las diferencias entre las puntuaciones debido a la falta de confiabilidad en sus respectivas puntuaciones.

Error estándar de la media Índice de variabilidad en las medias de las muestras alrededor de la media de la población.

Error estándar de medición Índice del grado de variabilidad en las puntuaciones de una prueba que resulta de una confiabilidad imperfecta.

Error no sistemático Error aleatorio e impredecible que se incorpora en la puntuación obtenida.

Escala de fingimiento de mala imagen Escala del MMPI-2 diseñada para detectar el falseamiento negativo.

Escala de razón Tipo de escala que clasifica y, luego, ordena los objetos a lo largo de un continuo con intervalos iguales y un verdadero punto cero.

Escala de valoración conductual Conjunto de preguntas o reactivos acerca de las conductas específicas de un niño (p. ej., orden, atención, actos agresivos), por lo común contestados por maestros, padres u otros cuidadores.

Escala de valoración gráfica Escala de valoración en la que las respuestas pueden marcarse en cualquier punto a lo largo del continuo entre dos polos; después, las marcas se convierten en una forma numérica.

Escala intervalar Escala que ordena los puntos de los datos en intervalos iguales, pero carece de un cero absoluto.

Escala nominal Tipo primitivo de escalas que sólo ubica los objetos en categorías separadas, sin referencia a diferencias cuantitativas.

Escala ordinal Escala que ubica los objetos en orden sin implicar distancias iguales entre los puntos a lo largo de la escala.

Escáner Máquina para convertir las respuestas marcadas en una hoja de respuestas en una forma eléctrica o electrónica.

ESEA Elementary and Secondary Education Act [Ley de Educación Primaria y Secundaria], conjunto de leyes federales relacionadas con las escuelas públicas.

Especificidad Capacidad de una prueba de *no* seleccionar individuos que *no* tienen

alguna característica.

Estadística descriptiva Rama de la estadística que se dedica a describir características de los datos crudos.

Estadística inferencial Rama de la estadística que se ocupa de hacer inferencias acerca de poblaciones enteras con base en el análisis de los datos de muestras.

Estandarización Puede referirse a tener instrucciones fijas para una prueba o al proceso de elaborar normas para una prueba.

Estandarizado Casi siempre se refiere al uso de condiciones uniformes para aplicar y calificar una prueba; también puede significar que una prueba tiene normas.

Estanina Sistema de puntuaciones estándar con una media de 5 y una desviación estándar de aproximadamente 2, diseñado para contener la distribución entera en el rango de 1–9, con intervalos iguales excepto en los extremos.

Estilo de respuesta Véase **Dirección de las respuestas**.

Estímulos ambiguos Estímulos de una prueba que ofrecen pocos indicios de cómo responder, lo que promueve respuestas variadas; se usan típicamente en las técnicas proyectivas.

Ética Estudio de lo que debe (o no debe) hacerse de acuerdo con la ética, la moral y los principios profesionales; principios que ayudan a definir la conducta apropiada dentro de una profesión particular.

ETS Educational Testing Service, ubicado en Princeton, NJ, editorial y creador importante de pruebas.

ETS Test Collection Colección electrónica que ofrece información básica de más de 25 000 pruebas.

Evaluación adaptada por la computadora (EAC) Método para evaluar en el que los reactivos presentados a un examinado son determinados por las respuestas a los reactivos anteriores.

Evaluación del desempeño Evaluación que implica las respuestas a estímulos parecidos a los de la vida real.

Falseamiento negativo Responder reactivos para proyectar una imagen desfavorable.

Falseamiento positivo Responder reactivos para proyectar una imagen favorable.

Fase de indagación Segunda fase de la evaluación con una técnica proyectiva en la que el examinador indaga acerca de las razones de las respuestas que el examinado dio en la primera fase.

Fase de respuestas Primera fase de la evaluación con una técnica proyectiva en la que el examinador simplemente registra las respuestas del examinado.

FERPA Family Educational Rights and Privacy Act [Ley de Derechos Educativos y Privacidad de la Familia] de 1994; también se conoce como la enmienda Buckley.

Flexibilidad cognitiva Capacidad para cambiar de conjuntos cognitivos con relativa facilidad.

Forense Relacionado con cuestiones legales. La psicología forense es la aplicación de los principios y métodos psicológicos en el contexto legal.

Formato Likert Formato de los reactivos de actitud en el que un examinado expresa su grado de acuerdo o desacuerdo con una afirmación.

Formulación del propósito Formulación del autor acerca de qué busca medir la prueba, por lo general, incluyendo el grupo meta.

Frecuencia empírica extrema Reactivos de una prueba, sobre todo en el campo de la personalidad, respondidos en la misma dirección por casi todos los examinados.

Frenología Teoría de una relación entre las formaciones craneales y la personalidad.

Función informativa del reactivo Índice, en la TRR, de la contribución de un reactivo individual a la información proporcionada por la puntuación de una prueba.

Funcionamiento diferencial del reactivo (FDR) Procedimientos para determinar si los reactivos de una prueba operan de manera diferenciada en diversos grupos de examinados (p. ej., género o grupos étnicos/raciales).

Funciones ejecutivas Funciones mentales relacionadas con planeación, valoración, juicio y manejo de otras capacidades mentales.

g Capacidad mental general que se supone subyace en las correlaciones considerables positivas entre muchas pruebas mentales; el primero en promoverla fue Charles Spearman.

Galeno Médico romano (aproximadamente 200 d. de C.) quien señaló el papel crucial del cerebro.

Galton Francis Galton, pionero inglés de las pruebas mentales, instigó el desarrollo de varias técnicas estadísticas tempranas e introdujo conceptos de la teoría de la evolución de Darwin en la psicología.

Gall Creador del concepto de frenología.

Generalización de la validez Desarrollar un resumen de toda la información relacionada con la validez de una prueba.

GI Forum Proceso judicial de *GI Forum contra TEA* (Texas Education Association) que examinó la legalidad del programa estatal de evaluación de Texas en vista de la considerable disparidad en las puntuaciones entre estudiantes pertenecientes a las mayorías y a las minorías.

GRE *Graduate Record Examinations* [Exámenes de Registro para Graduados]; existe un examen General y de varios Temas.

Griggs Demandante principal en el caso federal *Griggs contra Duke Power*, famoso caso sobre el uso de pruebas para contratar empleados.

Grupo de conveniencia Grupo obtenido porque está disponible oportunamente en vez de formarlo de acuerdo con un plan racional de muestreo.

Grupo normativo Cualquier grupo cuyo desempeño en una prueba se use como base para interpretar las puntuaciones de otros individuos.

Grupos superior e inferior Grupos con puntuaciones altas o bajas en una prueba total; se usan para analizar los reactivos individuales.

Guttman Louis Guttman, creador del escalamiento Guttman (análisis de escalograma) para la medición de actitudes.

Hacerse el enfermo Fingimiento negativo en reactivos o un conjunto de pruebas.

Halstead-Reitan Neuropsychological Battery [Batería Neuropsicológica Halstead-Reitan] Una de las baterías fijas de pruebas neuropsicológicas más usadas.

Heterocedasticidad Diferentes grados de dispersión en distintos puntos a lo largo de la línea de regresión que mejor se ajusta a los datos.

Heterogeneidad Diferencias excesivas entre individuos, sobre todo más grandes de lo normal.

Hipótesis proyectiva Suposición de que, cuando se presenta un estímulo ambiguo, la respuesta de una persona estará determinada por las características y la dinámica de la personalidad.

HIPPA Health Insurance Portability and Accountability Act [Ley de Responsabilidad y Portabilidad de los Seguros de Salud] de 1996 (P.L. 104–91), que tiene implicaciones importantes para el tratamiento de la información de los clientes.

Histograma de frecuencias Representación gráfica de una distribución de frecuencia mediante barras, por lo general verticales, que corresponden a la frecuencia de casos por cada puntuación o intervalo de puntuaciones.

Holland John Holland, autor del sistema RIASEC y de la Búsqueda Autodirigida.

Homocedasticidad Igual grado de dispersión en varios puntos a lo largo de la línea de regresión que mejor se ajusta a los datos.

Homogeneidad Diferencias entre los individuos que son menores de lo normal.

HTP Prueba Casa-Árbol-Persona, en la que se pide al examinado que dibuje una casa, un árbol y una persona.

IDEA Individuals with Disabilities Education Act [Ley de Educación para Individuos con Discapacidades], ley federal relacionada con la identificación y tratamiento de individuos con discapacidades, incluyendo dificultades de aprendizaje.

IEP Individualized education program [programa individualizado de educación], que debe ofrecerse a cada individuo identificado con una discapacidad.

Impacto adverso Cuando un procedimiento de selección resulta en tasas de selección diferentes para distintas clases protegidas.

Índice de deterioro Puntuación basada en cinco pruebas del Halstead-Reitan que indica la presencia de un déficit.

Índice de heredabilidad Índice del porcentaje de variabilidad en un rasgo atribuible a la genética en oposición al ambiente.

Informe interpretativo Informar el desempeño en la prueba con palabras habituales más que con puntuaciones numéricas.

Inteligencia cristalizada En varias teorías de la inteligencia, es la parte de la inteligencia que resulta de la acumulación de experiencias específicas de aprendizaje.

Inteligencia fluida En varias teorías de la inteligencia, es la parte de la inteligencia que supuestamente no depende de experiencias de aprendizaje altamente específicas.

Inteligencias múltiples (IM) Teoría de los siete (o más) tipos de inteligencia de Howard Gardner, como la inteligencia interpersonal y corporal-cinética.

Interacción Aplicada a la herencia y al ambiente, noción de que estos factores operan de un modo multiplicativo más que aditivo.

Interpretación con referencia a un criterio Interpretar el desempeño en una prueba en relación con algún criterio externo bien definido y no en relación con normas; contrasta con la interpretación con referencia a una norma.

Interpretación con referencia a una norma Interpretación de las puntuaciones de una prueba en relación con el modo en que los grupos en verdad se han desempeñado en la prueba; contrasta con la interpretación con referencia a un criterio.

Intervalos aparentemente iguales Método para elaborar escalas de actitud que intentan crear puntos en una escala psicológicamente equidistante.

Inventario Conjunto de reactivos. El término "inventario" se usa en la medición de la personalidad como equivalente de "prueba".

Inventario integral Prueba de personalidad que intenta medir todas las dimensiones importantes de la personalidad, ya sea normal o psicopatológica.

Jensen Arthur Jensen, principal defensor de un método de procesamiento de información para estudiar la inteligencia, con énfasis especial en el uso de tareas cognitivas elementales.

Jurisprudencia Ley basada en los precedentes asentados por los fallos de los tribunales.

Karraker Proceso judicial federal de EUA de *Karraker contra Rent-A-Center* relacionado con la cuestión de si el MMPI era una prueba médica.

KCIA *Kuder Career Interests Assessments*, inventario de intereses vocacionales de Kuder en el que se presentan informes en los que se empareja al examinado con personas que ya trabajan y tienen perfiles de intereses similares.

KDF *Prueba de Dibujo Cinético de la Familia*, en la que el examinado dibuja una familia haciendo algo.

KR-20 Fórmula no. 20 de Kuder-Richardson; índice de consistencia interna.

KR-21 Fórmula no. 21 de Kuder-Richardson; índice de consistencia interna que supone que todos los reactivos tienen valores equivalentes de dificultad.

Kuder G. Fredrick Kuder, autor de una serie de pruebas de intereses vocacionales y coautor de las fórmulas de confiabilidad de Kuder-Richardson.

Larry P. Principal demandante en *Larry P. contra Riles*, caso federal en California acerca del uso de una prueba de CI para su ubicación en una clase de educación especial.

Ley administrativa Ley que tiene su origen en una agencia de gobierno; a menudo se denomina regulaciones.

Ley de Derechos Civiles Las Leyes de Derechos Civiles de 1964 y 1991; leyes federales que intentan eliminar la discriminación por origen étnico o racial, género y religión.

Ley de Rehabilitación Ley de EUA de 1973 que exige adaptaciones en las instalaciones públicas para personas con discapacidades.

Ley Declaraciones acerca de lo que uno está obligado a hacer (o a no hacer) de acuerdo con los mandatos legales. Contrástese con la ética.

Ley estatutaria Ley que resulta de la acción de una legislatura; contrasta con la ley administrativa y la jurisprudencia.

Licencia Procedimiento legal para permitir a alguien que ejerza una profesión o practique un arte.

Likert Rensis Likert, creador del método Likert para la medición de actitudes.

Línea de regresión Línea que mejor se ajusta y muestra la relación entre los datos de dos variables.

Luria-Nebraska Neuropsychological Battery [Batería Neuropsicológica Luria-Nebraska] Una de las baterías fijas de pruebas neuropsicológicas que más se usan.

Manejo de impresiones Responder de manera deliberada los reactivos de una prueba para crear cierta imagen o impresión, sin importar los verdaderos sentimientos de uno sobre los reactivos.

Materiales de la prueba Conjunto de materiales diseñados para reseñar la prueba que, por lo general, incluye cuadernillo(s), instrucciones de aplicación y calificación, y un manual con información técnica.

MCMI *Millon Clinical Multiaxial Inventory*, uno de los varios inventarios Millon. La principal característica del MCMI es su esfuerzo por alinear sus escalas con el DSM-IV.

Media Promedio; medida de tendencia central.

Mediana Puntuación a la mitad de la distribución cuando las puntuaciones se ordenan numéricamente; medida de tendencia central.

Medidas de intereses vocacionales Pruebas que relacionan las preferencias, intereses y disposiciones de una persona con posibles carreras o puestos de trabajo.

Memoria de trabajo Constructo mental que designa la retención de múltiples elementos de información en el almacenamiento y su manipulación de maneras significativas en el corto plazo.

Memoria demorada Recordar material después de una demora de 20 a 30 min.

Memoria inmediata Recordar material después de un breve periodo (es decir, unos cuantos segundos).

Metaanálisis Conjunto de técnicas para combinar los resultados de varios estudios

empíricos.

Método de evaluaciones sumarias Nombre técnico del método de Likert para medir actitudes; implica sumar las evaluaciones de las personas a los reactivos presentados en formato Likert.

Método de muestreo de experiencias Muestreo sistemático de períodos de tiempo para determinar la ocurrencia de ciertas conductas.

MMPI *Inventario Multifásico de Personalidad de Minnesota*, medida de personalidad muy usada que se centra primordialmente en la psicopatología.

MMPI-RF "Forma reestructurada" del MMPI que usa un subconjunto de reactivos de dicha prueba e intenta rectificar varios de sus defectos.

MMY *Mental Measurements Yearbook(s)*, colecciones de reseñas de pruebas que se publican de manera periódica, primero a cargo de O. K. Buros y, ahora, del Buros Center for Testing; véase **Buros**.

Moda Puntuación que ocurre con la mayor frecuencia en una distribución de puntuaciones; medida de tendencia central.

Modelo biológico Modelo de la inteligencia que hace hincapié en las raíces biológicas de las operaciones mentales, en especial de las funciones cerebrales.

Modelo del procesamiento de información Cualquier modelo de la inteligencia que se concentra en cómo la mente trata los elementos de la información.

Modelo jerárquico Modelo de la inteligencia que postula un ordenamiento semejante a un árbol de las capacidades específicas que se agregan a capacidades superiores, sucesivamente más generales.

Modelo Rasch Modelo de un parámetro de la TRR; supone que los reactivos tienen valores iguales de discriminación y difieren sólo en el nivel de dificultad.

Modificación Cambios en el ambiente o en las condiciones de aplicación diseñados para eliminar los efectos de una discapacidad en el desempeño en la prueba; los cambios pueden ser suficientemente considerables para que la prueba, quizá, no mida ya el mismo constructo.

Monocigótico Gemelos que resultan de un óvulo –fertilizado por un espermatozoide–, que luego se divide en dos, por lo que los dos individuos tienen la misma dotación genética; contrasta con **dicigótico**.

NAEP National Assessment of Educational Progress, programa para investigar conocimientos y habilidades en varios dominios de contenido en EUA.

NCLB No Child Left Behind Act [Ley Que Ningún Niño se Quede Atrás] de 2001. Ley federal que exige a los estados elaborar y aplicar pruebas de aprovechamiento en muchos grados, con énfasis en llevar a todos los niños a un nivel de "competencia" en el desempeño.

Negativo falso Caso que excede un punto de corte en el criterio, pero que no excede el punto de corte de la prueba que intenta predecir el criterio.

NEO PI *NEO* (Neuroticismo, Extroversión, Franqueza) *Personality Inventory*, medida

muy usada de los Cinco Grandes rasgos de personalidad.

Neuropsicología clínica Especialidad profesional que combina la neuropsicología humana y la psicología clínica.

Neuropsicología Estudio de las relaciones cerebro-conducta.

Neutralidad de la prueba Grado en que una prueba mide determinado constructo de manera equivalente en distintos grupos. Es lo opuesto del sesgo de la prueba.

New Haven Proceso judicial de los bomberos de New Haven (CT), cuyo nombre oficial es *Ricci et al. contra DeStefano et al*., sobre la legalidad de usar información de las pruebas en vista de las considerables diferencias entre los blancos y los grupos minoritarios.

NOCTI National Occupational Competency Testing Institute.

Norma de desarrollo Norma de una prueba basada en el nivel de desarrollo del rasgo o característica que se mide.

Norma institucional Normas basadas en el desempeño promedio (o mediano) de instituciones enteras más que en individuos.

Norma local Norma basada en un grupo local de individuos; por lo general, contrasta con una norma nacional.

Norma nacional Norma basada en un grupo que, se pretende, sea representativo de la nación entera.

Normas del usuario Normas basadas en todos los casos que contestaron una prueba, al menos dentro de cierto periodo de tiempo específico.

Normas por subgrupo (obtención de normas) Elaboración de normas separadas para cada subgrupo, por ejemplo, diferentes grupos raciales/étnicos o de género, sobre todo con el fin de igualar las razones de selección.

Normas Resúmenes numéricos de las puntuaciones que obtuvieron las personas en el programa de estandarización de la prueba.

Observación conductual análoga Técnica para observar la conducta en una situación que simula la vida real; técnica de evaluación conductual.

OLSAT *Otis-Lennon School Ability Test*.

P.L. Abreviatura de Public Law [Ley Pública].

Parámetro de adivinación En un modelo de tres parámetros de la TRR, parámetro que estima las probabilidades de responder correctamente adivinando.

PASE Proceso judicial *PASE contra Hannon* sobre el uso de las pruebas de inteligencia para asignar a clases de educación especial. PASE significa Parents in Action on Special Education [Padres en acción en la educación especial].

Pendiente Inclinación de una línea que describe la relación entre dos variables; en la TRR, lo empinado de la curva característica del reactivo.

Percentil Punto en una escala debajo del cual cae el porcentaje especificado de casos.

Perseveración Tendencia a continuar haciendo lo mismo; incapacidad para cambiar los

patrones de pensamiento.

Perspectiva diferencial Disposición general para ver la conducta humana en términos de las diferencias entre las personas más que en términos de leyes generales que se aplican a todos.

Piers-Harris *Piers-Harris Children's Self-Concept Scale* [Escala de Autoconcepto Infantil de Piers-Harris].

Polígono de frecuencias Representación gráfica de una distribución de frecuencia mediante puntos unidos con una línea que corresponden a la frecuencia de casos por cada puntuación o intervalo de puntuaciones.

Populares Respuestas muy comunes o frecuentes ante un estímulo particular en una prueba proyectiva.

Positivo falso Caso que excede el punto de corte en una prueba que busca predecir un criterio, pero no excede el punto de corte del criterio.

PPVT *Peabody Picture Vocabulary Test*.

Práctica basada en la evidencia Prácticas médicas y psicológicas basadas en prácticas y procedimientos científicos sólidos, incluyendo las evaluaciones.

Precisión de la medición Índice de confiabilidad derivado de la teoría de la respuesta al reactivo que muestra qué tan bien se estimó una puntuación a partir del modelo y los reactivos.

Premórbido Tiempo previo al inicio del deterioro.

Procedimiento Mantel-Haenszel Técnica estadística para examinar las diferencias grupales en las respuestas a los reactivos individuales en comparación con el desempeño general en una prueba.

Procesamiento secuencial Procesamiento mental que avanza de una manera serial, siguiendo pasos.

Procesamiento simultáneo Procesamiento mental que opera sobre distintas fuentes de información casi al mismo tiempo.

Proceso de respuesta Procesos, por ejemplo, operaciones mentales, que una persona usa para responder los reactivos o terminar los ejercicios de una prueba.

Producción convergente Operaciones mentales que requieren que una persona desarrolle una respuesta correcta a un problema.

Producción divergente Operaciones mentales que requieren que una persona desarrolle muchas respuestas diferentes a un problema, sobre todo que sean novedosas o únicas.

Programa de estandarización Programa de investigación usado para establecer las normas de una prueba.

Protocolo Registro de las respuestas a una prueba.

Prueba culturalmente neutral Prueba que intenta ser igualmente neutral para individuos de diferentes entornos culturales.

Prueba de dominio específico Prueba que se enfoca en sólo una o algunas variables

del dominio no cognitivo; contrasta con los inventarios integrales.

Prueba de ejecución Prueba que requiere que las personas realicen cierta acción, sobre todo una acción que semeja una situación de la vida real; por lo general, contrasta con una prueba de lápiz y papel.

Prueba de ensayo Prueba que requiere escribir un ensayo en respuesta a una pregunta o indicación; de manera más general, se aplica a cualquier tipo de prueba distinta de las de opción múltiple.

Prueba de lápiz y papel Pruebas en las que los reactivos se presentan de manera escrita y las respuestas se registran del mismo modo, sobre todo marcando una hoja de respuestas; por lo general, contrasta con una prueba de ejecución.

Prueba de niveles múltiples Series de pruebas coordinadas a lo largo de edades o grados sucesivos.

Prueba de poder Prueba que demanda el máximo desempeño, por lo general con pocas restricciones de tiempo o ninguna.

Prueba de velocidad Prueba con un material relativamente fácil que se debe terminar tan rápido como sea posible.

Prueba estructurada Prueba, sobre todo en el dominio de la personalidad, con modos fijos de respuesta; suele contrastarse con las técnicas proyectivas, las cuales usan un formato de respuesta libre.

Prueba grupal Cualquier prueba susceptible de aplicarse a grupos grandes de manera simultánea.

Prueba individual Cualquier prueba psicológica diseñada para aplicarse a un solo individuo a la vez.

Prueba objetiva de personalidad Prueba de personalidad que se puede calificar de una manera sencilla, como si fuera un trabajo de oficina (p. ej., contando las respuestas a reactivos de opción múltiple o de falso o verdadero).

Prueba Proceso o dispositivo estandarizado que produce información acerca de una muestra de conducta o procesos cognitivos de una manera cuantificada.

Prueba situacional Prueba en que una persona se coloca en una simulación de una situación de la vida real para que se pueda observar su conducta.

Pruebas de aprovechamiento Pruebas diseñadas para medir conocimientos o habilidades, en especial las que se desarrollan a lo largo de las experiencias escolares y laborales.

Pruebas de capacidad mental Pruebas que se ocupan primordialmente de la inteligencia y capacidades relacionadas.

Pruebas de gran importancia Pruebas que tienen consecuencias importantes para los examinados (u otros individuos) como las pruebas de certificación o licencia.

Pruebas neuropsicológicas Pruebas diseñadas para medir las funciones del cerebro y del sistema nervioso.

Psicología positiva Movimiento dentro de la psicología que hace hincapié en los rasgos

positivos, las virtudes y las fortalezas del carácter, en oposición al énfasis en los padecimientos patológicos.

PsycTESTS Colección electrónica de pruebas financiada por la American Psychological Association.

Puntuación de corte Puntuación de una prueba o criterio que indica aprobación o fracaso, o alguna otra división; también se llama punto de corte.

Puntuación de Déficit Neuropsicológico General Puntuación derivada de muchas variables del Halstead-Reitan que indica la gravedad general del déficit neurológico.

Puntuación de error Diferencia hipotética entre la puntuación obtenida y la puntuación verdadera de una persona.

Puntuación de porcentaje de correctos Expresa el desempeño en la prueba como el porcentaje de reactivos contestados correctamente o en una dirección respecto del número total de reactivos de la prueba.

Puntuación de universo En la teoría de la generalizabilidad, puntuación que resulta de la aplicación teórica de una prueba en múltiples ocasiones, condiciones y muestras de contenido.

Puntuación estándar Tipo de norma en la que las puntuaciones naturales se convierten en una escala con una media y una desviación estándar nuevas; ambas se suelen seleccionar para ser números más amables.

Puntuación índice Puntuación basada en factores de cada una de las escalas Wechsler de inteligencia recientes.

Puntuación natural Resultado original de la calificación de una prueba, por ejemplo, número de respuestas correctas o en una misma dirección, antes de traducirlas a algún tipo de norma.

Puntuación normativa Cualquier puntuación interpretada en el marco de un conjunto de normas.

Puntuación observada Puntuación real de una persona en una prueba.

Puntuación T Sistema de puntuaciones estándar con $M = 50$ y $DE = 10$; a veces se denomina puntuación T de McCall.

Puntuación verdadera Puntuación que una persona obtendría teóricamente si todas las fuentes de varianza no confiable se eliminaran o cancelaran.

Puntuación z Puntuación que resulta de restar la media de una puntuación natural y, luego, dividir el resultado entre la desviación estándar, así $z = (X - M) / DE$; a veces se denomina puntuación desviada normal.

Puntuaciones escalares Suelen referirse a un tipo de puntuación estándar empleada para ligar varios niveles de una prueba de niveles múltiples.

Puntuaciones especiales Diversas puntuaciones muy focalizadas del Sistema Integral de Exner del Rorschach.

Puntuaciones estándar normalizadas Puntuaciones estándar que se obtienen convirtiendo una distribución no normal en una normal por medio de una

transformación no lineal de las puntuaciones.

Puntuaciones ipsativas Puntuación relativa a otras puntuaciones de un individuo.

Rango Distancia de la puntuación más baja a la más alta en un conjunto de datos.

Rango intercuartil Distancia del percentil 25 al 75 en una distribución; medida de variabilidad.

Rango percentil Porcentaje de casos del grupo de estandarización que cae debajo de determinada puntuación natural.

Rapport Atmósfera cálida, amigable, sobre todo la que se establece al inicio de una sesión de evaluación.

Raven Matrices Progresivas de Raven (en cualquiera de sus versiones).

Reactivo de respuesta abierta Reactivo de una prueba que demanda que el examinado construya una respuesta en vez de elegirla a partir de ciertas alternativas.

Reactivo tipo matriz Reactivo de una prueba que presenta una matriz con algún tipo de patrón que el examinado debe completar.

Reactivos de rellenar casillas Tipo de reactivos, por lo general numéricos, en que el examinado responde rellenando casillas.

Reactivos de respuesta cerrada Reactivos en los que el examinado elige una respuesta de las alternativas que se le presentan.

Regla de cuatro quintos Regla que aparece en las directrices EEOC que operacionaliza la definición de impacto adverso como una diferencia de 80% entre grupos con las tasas de selección más alta y más baja.

Reglas de inicio y discontinuación En una prueba de aplicación individual que cubre un amplio rango de capacidades, reglas de dónde empezar y dónde terminar con un individuo en el rango entero de reactivos.

Regulaciones Término alternativo de la ley administrativa, como en el Código de Regulaciones Federales (Code of Federal Regulations).

Replicabilidad Término técnico en el escalamiento Guttman que designa qué tan bien los patrones de respuesta de las personas se ajustan al modelo teórico de una escala Guttman.

Requisitos que debe cubrir el usuario de pruebas Credenciales necesarias para comprar pruebas, a menudo representadas en tres niveles.

Responsabilidad Movimiento en la educación que exige a las escuelas y funcionarios públicos demostrar el éxito de los programas educativos, a menudo con ayuda de pruebas.

Respuestas socialmente deseables Respuestas a los reactivos que van en dirección de lo socialmente deseable o de lo "políticamente correcto", en especial cuando no son consistentes con los verdaderos sentimientos del examinado.

Restricción de rango Variabilidad reducida en una o más variables, en especial la que afecta la correlación entre ellas; también se denomina atenuación.

Resumen estructural Resumen general de muchas puntuaciones que resultan de la aplicación del Sistema Integral de Exner del Rorschach.

Retención de dígitos Prueba que implica la memoria a corto plazo en una serie de dígitos aleatorios.

Revisión de panel Procedimiento en el que se cuenta con un panel de representantes de grupos minoritarios que revisan los reactivos de la prueba para detectar los que pueden poner en desventaja a los miembros del grupo minoritario.

RIASEC Iniciales de Realista, Investigador, Artístico, Social, Emprendedor y Convencional en el hexágono de Holland que representa los tipos de personalidad y los ambientes de trabajo.

RISB *Rotter Incomplete Sentences Blank* [Frases Incompletas de Rotter].

Role-playing Adoptar cierto papel, sobre todo con propósitos de evaluación o terapia.

Romper el hielo Uso de una prueba para empezar la conversación entre examinador y examinado, en especial para hacer sentir cómodo al examinado.

Rorschach Se refiere a Hermann Rorschach, quien experimentó ampliamente con manchas de tinta como estímulos de una prueba, y a las propias manchas de tintas, con el método asociado de presentación.

SAT Suele referirse al SAT I: Prueba de Razonamiento, antes llamada Scholastic Assessment Test [Prueba de Aprovechamiento Académico] o Scholastic Aptitude Test [Prueba de Aptitud Académica]; a veces se refiere al Stanford Achievement Test [Prueba de Aprovechamiento de Stanford].

SCID *Structured Clinical Interview for DSM-IV* [Entrevista Clínica Estructurada para el DSM-IV], la mejor entrevista clínica estructurada conocida que busca producir un diagnóstico de acuerdo con la clasificación del DSM.

SCL-90 R *Symptom Checklist 90, Revised* [Lista de Revisión de Síntomas, Revisada], medida relativamente breve (90 reactivos) de síntomas clínicos.

SDS *Self-Directed Search*, prueba de Holland de intereses vocacionales.

Seguridad de la prueba Mantener los materiales de la prueba en lugares donde sólo personas con los conocimientos y habilidades apropiadas puedan tener acceso.

Sensibilidad Exactitud, en forma de porcentaje, con la que una prueba identifica individuos con cierta característica.

Sesgo de la pendiente Sesgo de la prueba que se demuestra cuando las líneas de regresión de dos grupos tienen distintas pendientes.

Sesgo de la pendiente Sesgo de la prueba que se presenta cuando las líneas de regresión de los dos grupos tienen pendientes diferentes.

Sesgo de la prueba Mostrar que una prueba mide constructos un poco diferentes en distintos grupos de examinados, sobre todo en grupos mayoritarios y minoritarios.

Seudodemencia Deterioro cognitivo similar a la demencia tipo Alzheimer, pero que es resultado de un padecimiento psiquiátrico, por lo general depresión.

SII *Strong Interest Inventory*, versión actual del antiguo *Strong Vocational Interest Blank* (SVIB) y del *Strong-Campbell Interest Inventory* (SCII).

Sistema de puntos Método para calificar pruebas en donde se conceden puntos por reactivo o por cada parte del reactivo.

Sistema integral Nombre del método de Exner para aplicar y codificar (calificar) el Rorschach.

Spearman Charles Spearman, creador inglés de la teoría de "*g*", inteligencia general; también desarrolló las primeras formas del análisis factorial.

STAI *State Trait Anxiety Inventory* [Inventario de Ansiedad Rasgo-Estado]

Stanford-Binet *Stanford-Binet Intelligence Scale* [Escala de Inteligencia Stanford-Binet], sucesora estadounidense de la escala original de Alfred Binet.

Strong Edward K- Strong, pionero en la medición de intereses vocacionales y principal autor de diversos inventarios Strong.

Sub-representación del constructo Fracaso para medir por completo el constructo que queremos calcular; medir sólo parte del constructo de interés.

Tamaño del efecto Medida de la magnitud de un fenómeno estadístico, en especial de uno independiente del tamaño de la muestra.

Tareas cognitivas elementales Tareas relativamente sencillas que se usan para estudiar las operaciones y capacidades mentales.

Tasa base Tasa en la que algunas características aparecen en una población.

TAT *Test de Apercepción Temática*.

Técnica de hablar en voz alta Decir las reacciones, pensamientos y sentimientos mientras se vive alguna experiencia, con énfasis especial en el informe detallado.

Técnicas proyectivas Método de evaluación en el que los estímulos son relativamente ambiguos y la persona tiene una libertad considerable para responder.

Temas de diseño Decisiones pendientes acerca del diseño de una prueba, como su extensión, formato de respuesta, número de puntuaciones, etc.

Tendencia central Estadístico que describe las puntuaciones medias o intermedias de una distribución; las medidas usuales de tendencia central son media, mediana y moda.

Teoría bifactorial Teoría de las capacidades mentales generales y específicas de Spearman.

Teoría clásica de las pruebas Teoría tradicional acerca de la construcción y confiabilidad de las pruebas que incorpora la teoría de la puntuación verdadera.

Teoría de factores múltiples Cualquier teoría de la inteligencia que hace hincapié en más de una dimensión; sobre todo teorías que surgen de los métodos analítico-factoriales como las capacidades mentales primarias de Thurstone.

Teoría de la generalizabilidad Método para estudiar la confiabilidad de las pruebas que permite examinar varias fuentes de varianza no confiable al mismo tiempo.

Teoría de la respuesta al reactivo (TRR) Método de construcción de prueba que usa la curva característica del reactivo.

Teoría de los tres estratos Modelo jerárquico de Carroll, con numerosas capacidades específicas, ocho factores de nivel intermedio y coronados con "*g*".

Teoría PASS Teoría de la inteligencia construida alrededor de la planeación, atención y procesamiento simultáneo y secuencial.

Teoría triárquica Teoría de la inteligencia de Sternberg que postula tres subteorías: componencial, experiencial y contextual.

Teorías basadas en etapas Teorías que hacen hincapié en el desarrollo mediante la progresión a través de etapas cualitativamente distintas.

Teorías del desarrollo Teorías de la inteligencia que hacen hincapié en que existen etapas en el desarrollo de las operaciones y capacidades mentales, sobre todo en el paso de una etapa a otra.

Teorías psicométricas Teorías de la inteligencia que dependen en gran parte del uso de pruebas y el examen de la relación entre ellas.

Test Critiques Colección de reseñas de pruebas publicada por PRO-ED, Inc.

Tests Lista exhaustiva de pruebas con información descriptiva básica, publicada por PROD-ED, Inc.

Theta Puntuación derivada de la aplicación de la teoría de la respuesta al reactivo en el desempeño en la prueba.

Thurstone Louis Thurstone, creador de la teoría de las capacidades mentales primarias y contribuidor importante al desarrollo de los métodos analítico-factoriales y el escalamiento de actitudes.

TIMSS Trends in International Mathematics and Science Study [Tendencias en el Estudio Internacional en Matemáticas y Ciencia], programa para la evaluación de matemáticas y ciencia en muchos países.

TIP *Tests in Print*, serie de publicaciones que proporciona una lista exhaustiva de pruebas, publicada por el Buros Institute.

Tipo de código Código de dos o tres dígitos que indica los números de la escala de las puntuaciones más altas en un perfil de algunas pruebas como el MMPI y algunas encuestas de intereses vocacionales.

Transformación lineal Transformación de puntuaciones naturales de una escala original en una nueva escala que preserva las características de la original, a excepción de sus valores numéricos.

Transformación no lineal Transformación de las puntuaciones naturales que cambia las distancias entre los valores de la escala original a la nueva.

Tronco del reactivo Pregunta o parte de apertura de un reactivo al que el examinado debe proveer de una respuesta.

USC Abreviatura de U.S. Code [Código de EUA], es decir, ley estatutaria federal.

VABS *Vineland Adaptive Behavior Scales* [Escalas de Conducta Adaptativa de Vineland].

Validación cruzada Después de terminar un estudio de validación con un grupo, en especial donde los reactivos o variables se han seleccionado, realizar otro estudio de validación con un grupo diferente.

Validez aparente Apariencia de que una prueba mide el constructo meta, en especial cuando no está respaldada por ninguna otra evidencia empírica.

Validez concurrente Validez de la prueba que se demuestra mediante la relación entre ella y algún otro criterio medido aproximadamente al mismo tiempo.

Validez consecuencial Validez de la prueba que se define por las consecuencias de su uso con un propósito particular.

Validez convergente Evidencia de la validez que muestra que el desempeño en una prueba concuerda con otras medidas del constructo meta.

Validez de constructo Amplio conjunto de métodos empleados para apoyar la idea de que una prueba mide su constructo meta.

Validez de contenido Validez de la prueba que se define por la correspondencia entre el contenido de la prueba y algún otro cuerpo bien definido de materiales, como un currículum o un conjunto de habilidades para un trabajo.

Validez de criterio Demostrar la validez de una prueba mostrando la relación entre sus puntuaciones y algún criterio externo.

Validez discriminante Evidencia de la validez que muestra que el desempeño en una prueba tiene una correlación relativamente baja con medidas de constructos de los que se esperan correlaciones bajas con el rasgo de interés.

Validez incremental Aumento en la validez alcanzada agregando una nueva prueba o procedimiento a los existentes.

Validez Indicación del grado en que una prueba mide lo que pretende medir para un propósito específico.

Validez instruccional Demostración de que las personas que contestan una prueba de aprovechamiento estuvieron expuestas al material o tuvieron la oportunidad de aprenderlo.

Validez predictiva Validez que se demuestra mostrando el grado en que una prueba puede predecir el desempeño en algún criterio externo cuando se aplica por adelantado.

Valor p Porcentaje de reactivos correctos o contestados en cierta dirección en una prueba.

Variabilidad Grado de dispersión o de diferencia entre las puntuaciones de un conjunto de datos.

Variable Constructo o dimensión a lo largo de la cual los objetos varían.

Varianza compartida por la familia En estudios de herencia y ambiente, varianza atribuible al hecho de que los miembros de una familia presumiblemente tienen

ambientes similares.

Varianza Cuadrado de la desviación estándar; medida de variabilidad.

Varianza irrelevante al constructo Varianza en las puntuaciones de las pruebas asociada con variables distintas de las que queremos medir.

Vineland Término que se usa para designar al *Vineland Social Maturity Scale* o a su sucesor, el *Vineland Adaptive Behavior Scales*.

WAIS *Escala Wechsler de Inteligencia para Adultos*.

Wernicke Neuroanatomista alemán que encontró que el trastorno del lenguaje implicaba una comprensión deteriorada, pero con el habla intacta, aunque no era significativa; también designa el área del cerebro involucrada.

WISC *Escala Wechsler de Inteligencia para Niños*.

WMS *Wechsler Memory Scale* [Escala de Memoria de Wechsler].

WPPSI *Escala Wechsler de Inteligencia para Preescolar y Primaria*.

REFERENCIAS

AAIDD Ad Hoc Committee on Terminology and Classification. (2010). *Intellectual disability: Definition, classification, and systems of supports* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Achenbach, T. M.,&Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms and Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Ackerman, T. (1998). Review of the Wechsler Individual Achievement Test. In J. C. Impara & B. S. Plake (Eds.), *The thirteenth mental measurements yearbook* (pp. 1125– 1128). Lincoln, NE: Buros Institute of Mental Measurements.

Acklin,M.W. (1996). Personality assessment and managed care. *Journal of Personality Assessment, 66,* 194–201.

ACT. (2001). ACT assessment: Sample questions. Retrieved October 5, 2001, from http://www.actrs8.act.org/aap/sampletest/

ACT. (2007). ACT technical manual. Iowa City, IA: Author.

Adams,K.M. (1980). In search of Luria's battery:Afalse start. *Journal of Consulting and Clinical Psychology, 48,* 511–516.

Adams, K. M. (1984). Luria left in the lurch: Unfulfilled promises are not valid tests. *Journal of Clinical Neuropsychology, 6,* 455–465.

Aiken, L. R. (1999). *Personality assessment methods and practices* (3rd ed.). Seattle: Hogrefe & Huber.

Alfano, D. P.,Neilson, P.M., Paniak, C. E.,&Finlayson, M. A. J. (1992). TheMMPI and closed head injury. *The Clinical Neuropsychologist, 6,* 134–42.

Allison, D. B., & Baskin, M. L. (Eds.). (2009). *Handbook of assessment methods for eating behaviors and weight-related problems:Measures, theory, and research* (2nd ed.).Thousand Oaks, CA: Sage.

American Association on Mental Retardation. (2002). *Mental retardation: Definition, classification, and systems of support* (10th ed.). Washington, DC: Author.

American Counseling Association. (2000). *ACA code of ethics and standards of practice*. Alexandria, VA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1955). *Technical recommendations for achievement tests.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2013). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.) Text Revision. Washington, DC: Author.

American Psychiatric Association. (2002). The American Psychiatric Association's resource document on

mental retardation and capital sentencing: Implementing Atkins v. Virginia. *Journal of the American Academy of Psychiatry and the Law, 32,* 304–308.
American Psychiatric Association. (2013). Diagnostic *and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.
American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.
American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.
American Psychological Association. (1992). *Ethical principles of psychologists and code of conduct.* Washington, DC: Author.
American Psychological Association. (1994). Guidelines for child-custody evaluations in divorce proceedings. *American Psychologist, 49,* 677–680.
American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct.* Washington, DC: Author.
American Psychological Association. (2012). Guidelines for the evaluation of dementia and cognitive change. *American Psychologist, 67,* 1–9.
American Psychological Association. (2012). Guidelines for assessment of and intervention with persons with disabilities. American Psychologist, 67(1), 43–62.
Anastasi,A. (1954). *Psychological testing.* NewYork: Macmillan.
Andrews, J. J. W., Saklofske, D. H., & Janzen, H. L. (Eds.). (2001). *Handbook of psychoeducational assessment: Ability, achievement, and behavior in children.* San Diego, CA: Academic Press.
Ang, R. P., Lowe, P. A., & Yusof, N. (2011). An examination of RCMAS-2 scores across gender, ethnic background, and age in a large Asian school sample. *Psychological Assessment, 23,* 899–910.
Anholt, R. R. H., & Mackay, T. F. C. (2010). *Principles of behavioral genetics.* London, UK: Elsevier.
Arbisi, P. A. (2001). Review of the Beck Depression Inventory-II. In B. S. Plake & J. C. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 121–123). Lincoln: University of Nebraska Press.
Archer, R. P. (1992). Review of the Minnesota Multiphasic Personality Inventory-II. In J. J. Kramer & J. C. Conoley (Eds.), *Eleventh mental measurements yearbook* (pp. 546–562). Lincoln: University of Nebraska Press.
Archer,R. P.,Maruish, M., Imhof, E. A., & Piotrowski, C. (1991). Psychological test usage with adolescents: 1990 survey findings. *Professional Psychology: Research and Practice, 22,* 241–252.
Aristotle (1935). On *the soul. Parvanaturalia. On breath* (W.S. Hett,Trans.). Cambridge, MA: HarvardUniversityPress.
Ash, P. (1995). Review of the Eating Disorder Inventory- 2. In J. C. Conoley & J. C. Impara (Eds.), *Twelfth mental measurements yearbook* (pp. 333–335). Lincoln: University of Nebraska Press.
ASVAB Career Exploration Program. (2010). Theoretical and technical underpinnings of the revised skill composites and OCCU-Find. Accessed from http://www.asvabprogram.com/
Atkins V. Virginia, 536 U.S. 304 (2002).
Atkinson, J.W. (Ed.). (1958). *Motives in fantasy, action, and society.* Princeton, NJ: D. Van Nostrand.
Atkinson, L. (1986). *The comparative validities of the Rorschach and MMPI: A meta-analysis.* Canadian Psychology, 27, 238–247.
Atkinson, L., Quarrinton, B., Alp, I. E., & Cyr, J. J. (1986). Rorschach validity: An empirical approach to the literature. *Journal of Clinical Psychology, 42,* 360–362.
Baer, L., & Blais,M. A. (Eds.). (2010). *Handbook of clinical rating scales and assessment in psychiatry and mental health.* New York: Humana Press.
Baker, D. B., & Benjamin, L. T., Jr. (2000). The affirmation of the scientist-practitioner. *American Psychologist, 55(2),* 241–247.
Baker, F. B. (1971). Automation of test scoring, reporting, and analysis. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 202–234). Washington, DC: American Council on Education.
Baker, F. B. (1993). Computer technology in test construction and processing. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 409–428). Phoenix, AZ: American Council on Education and The Oryx Press.
Barker, C., Pistrang, N., & Elliott, R. (1994). *Research methods in clinical and counseling psychology.* New

York: Wiley.
Barrouillet, P., & Gaillard, V. (Eds.). (2011). *Cognitive development and working memory: A dialogue between neo- Piagetian and cognitive approaches.* New York: Psychology Press.
Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI–II manual.* San Antonio, TX: The Psychological Corporation.
Beck, S. J. (1937). Introduction to the Rorschach method: A manual of personality study. American Orthopsychiatric Association Monograph, no. 1, xvþ278.
Bellak, L. (1993). *The Thematic Apperception Test, the Children's Apperception Test and the Senior Apperception Test in clinical use* (5th ed.). Boston: Allyn Bacon.
Bellak, L., & Abrams, D. M. (1997). *The Thematic Apperception Test, the Children's Apperception Test, and the Senior Apperception Test in clinical use* (6th ed.). Boston: Allyn and Bacon.
Belter, R.W.,&Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology, 57,* 717–726.
Ben-Porath, Y. S. (1997). Use of personality assessment instruments in empirically guided treatment planning. *Psychological Assessment, 9,* 361–367.
Ben-Porath, Y. S., & Tellegen, A. (2008a). MMPI-2-RF (Minnesota Multiphasic Inventory-2-Restructured Form) Manual for administration. Minneapolis, MN: University of Minnesota.
Ben-Porath, Y. S., & Tellegen, A. (2008b). MMPI-2- RF (Minnesota Multiphasic Inventory-2-Restructured Form) User's guide for reports. Minneapolis, MN: University of Minnesota.
Benton, A. (1987). Evolution of a clinical specialty. *The Clinical Neuropsychologist, 1,* 5–8.
Benton, A. (1997). On the history of neuropsychology: An interview with Arthur Benton, Ph. D. *Division of Clinical Neuropsychology Newsletter, 40,* 15(2), 1–2; 14–16.
Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press.
Berk, R. A. (Ed.). (1984). *A guide to criterion-referenced test construction.* Baltimore: Johns Hopkins University Press.
Bessai, F. (2001). Review of the Peabody Picture Vocabulary Test-III. In B. S. Plake & J. C. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 908–909). Lincoln: University of Nebraska Press.
Bessette, J. M. (Ed.). (1996). American justice. Pasadena, CA: Salem Press.
Betsworth, D. G., & Fouad, N. A. (1997). Vocational interests: A look at the past 70 years and a glance at the future. The *Career Development Quarterly, 46,* 23–47.
Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique, 12,* 191–244 (translation by Elizabeth S. Kite, 1916, in *The development of intelligence in children.* Vineland, NJ: Publications of the Training School at Vineland).
Binet, A.,& Simon, T. (1916). *The development of intelligence in children* (E. S. Kite, Trans.) Baltimore: Williams and Wilkins.
Binet, A., & Simon, T. (with Terman, L. M.) (1980). *The development of intelligence in children.* Nashville, TN: Williams Printing Company.
Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, handbook I: Cognitive domain.* New York: Longman.
Board of Trustees of the Society for Personality Assessment. (2006). Standards for education and training in psychological assessment: Position of the Society for Personality Assessment. *Journal of Personality Assessment, 87*(3), 355–357.
Boring, E. D. (1950). *A history of experimental psychology* (2nd ed.). New York: Appleton-Century-Crofts.
Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48,* 299–337.
Botwin, M. D. (1995). Review of the Revised NEO Personality Inventory. In J. C. Conoley & J. C. Impara (Eds.), *Twelfth mental measurements yearbook* (pp. 861–863). Lincoln: University of Nebraska Press.
Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist, 44*(3), 576–578.
Boyle, G. J. (1995). Review of the Rotter Incomplete Sentences Blank, Second Edition. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 880–882). Lincoln: University of

Nebraska Press.
Bracken, B. A. (1992). *Multidimensional Self Concept Scale, Examiner's Manual.* Austin, TX: PRO-ED.
Breland, H.M., Kubota, M. Y., Nickerson, K., Trapani, C. & Walker, M. (2004). *New SAT writing prompt study: Analyses of group impact and reliability.* New York: College Board.
Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*(1), 5–10.
Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement,* 38, 295–317.
Brennan, R. L. (2001b). *Generalizability theory.* New York: Springer-Verlag.
Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24,* 1–21.
Bridgeman, B., Burton, N., & Cline, F. (2008). Understanding What the Numbers Mean: A Straightforward Approach to GRE Predictive Validity: ETS RR_08-06. Princeton, NJ: Educational Testing Service.
Bridgeman, B., Mccamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning Test,* Research report No. 2000-1. New York: College Entrance Examination Board.
Brim, O. G., Crutchfield, R. S., & Holtzman, W. H. (Eds.). (1966). *Intelligence: Perspectives 1965*. New York: Harcourt, Brace & World.
Brodsky, S. L. (1991). *Testifying in court: Guidelines and maxims for the expert witness.* Washington, DC: American Psychological Association.
Brodsky, S. L. (1999). *The expert witness: More maxims and guidelines for testifying in court.* Washington, DC: American Psychological Association.
Brody, N. (1992). *Intelligence* (2nd ed.). San Diego: Academic Press.
Brookings, J. B. (1994). Eating Disorder Inventory-2. In D. J. Keyser & R. C. Sweetland (Eds.), *Test Critiques,* vol. X (pp. 226–233). Kansas City, MO: Test Corporation of America.
Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. GRIMM & F. R. YARNOLD, *Reading and understanding multivariate statistics* (pp. 99–136). Washington, DC: American Psychological Association.
Bubenzer, D. L., Zimpfer, D. G., & Mahrle, C. L. (1990). Standardized individual appraisal in agency and private practice: A survey. *Journal of Mental Health Counseling, 12,* 51–66.
Buck, J. N. (1948). The H-T-P technique, a qualitative and quantitative scoring manual. *Journal of Clinical Psychology, 4,* 317–396.
Buck, J. N. (1966). *The House-Tree-Person Technique, revised manual.* Los Angeles: Western Psychological Services.
Burns, R. C., & Kaufman, S. H. (1970). *Kinetic Family Drawings (KFD): An introduction to understanding children through kinetic drawings.* New York: Brunner/Mazel.
Burns, R.C.,&Kaufman, S. H. (1972). *Action, styles, and symbols in Kinetic Family Drawings (KFD).* New York: Brunner/Mazel.
Burton, N. W., & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980.* Research report No. 2001–2. New York: College Entrance Examination Board.
Burton, N. W., & Wang, M. (2005). *Predicting long-term success in graduate school: A collaborative validity study* (GRE Board Report No, 99-14R, ETS RR-05-03). Princeton, NJ: Educational Testing Service.
Butcher, J. N. (1993). *Minnesota Multiphasic Personality Inventorv-2 (MMPI-2) User's guide for The Minnesota Report: Adult Clinical System—Revised.* Minneapolis: University of Minnesota Press.
Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2) manual for administration and scoring.* Minneapolis: University of Minnesota Press.
Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 content scales.* Minneapolis: University of Minnesota Press.
Camara, W. J. (2001). Do accommodations improve or hinder psychometric qualities of assessment? *The Score Newsletter, 23*(4), 4–6.
Camara, W., Nathan, J., & Puente, A. (1998). *Psychological test usage in professional psychology: Report to*

the APA practice and sciences directorate. Washington, DC: American Psychological Association.
Camara, W., Nathan, J., & Puente, A. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*(2), 141–154.
Campbell, D. T., & Fiske, D. W. (1954). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.
Cannon, B. J. (2000). A comparison of self- and otherrated forms of the Neuropsychology Behavior and Affect Profile in a traumatic brain injury population. *Archives of Clinical Neuropsychology, 15,* 327–334.
Carroll, J. B. (1993). *Human cognitive abilities: Asurvey of factor analytic studies.* New York: Cambridge University Press.
Carver, C. S., & Scheier, M. F. (1996). *Perspectives on personality* (3rd ed.). Needham Heights, MA: Allyn & Bacon.
Cattell, J. M. (1890). *Mental tests and measurements.* Mind, 15, 373–381.
Cattell, R. B. (1940). A culture-free intelligence test I. *Journal of Educational Psychology, 31,* 161–179.
Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.
Cattell, R. B. (1966). *The scientific analysis of personality.* Chicago: Aldine.
Chaplin, W. F. (1984). State-Trait Anxiety Inventory. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques,* vol. I (pp. 626–632). Kansas City, MO: Test Corporation of America.
Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology, 130*(3), 290–304.
Cheng, L. (1994).A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18,* 205–215.
Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9,* 233–255.
Chinn, R. N., & Hertz, N. R. (2002). Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education, 15*(1), 1–14.
Choca, J. P. (2001). Review of the Millon Clinical Multiaxial Inventory—III. In B. S. Plake &J. C. Impara (Eds.), *The fourteenth mental measurements yearbook* (pp. 765–767). Lincoln: University of Nebraska Press.
Christensen, A. L. (1984). *Luria's neuropsychological investigation.* Copenhagen, Denmark: Monksgaard.
Cizek, G. J. (2001a). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27.
Cizek, G. J. (2003). Review of the Woodcock-Johnson(r) III. In: Plake, B. S., Impara, J. C., & Spies, R. A. (Eds.). *The fifteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
Cizek, G. J. (Ed.). (2001b). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.
Cizek, G. J., & BUNCH, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.
Cizek, G. J.,Bowen, D.,&Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement, 70*(5), 732–743.
Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50.
Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68,* 397–412.
Clarke, A. M., & Clarke, A. D. (1985). Criteria and classification. In A. M. Clarke, A. D. Clarke, & J. M. Berg (Eds.), *Mental deficiency: The changing outlook* (4th ed., pp. 27–52). New York: Free Press.
Clauser, B. E., Swanson, D. B.,&Clyman, S. G. (1999). A comparison of the generalizability of scores produced by expert raters and automated scoring systems. *Applied Measurement in Education, 12,* 281–299.
Clemence, A. J., & Handler, L. (2001). Psychological assessment on internship: A survey of training directors and their expectations for students. *Journal of Personality Assessment, 76,* 18–47.
Cole, N. S.,&Moss, P. A. (1993). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.

201– 219). Phoenix, AZ: Oryx Press.
College Board. (2012). Test characteristics of the SAT: Reliability, difficulty levels, completion rates. Accessed at http://media.collegeboard.com/digital Services/ pdf/research/Test-Characteristics-of % 20-SAT-2012.pdf, 1-10-2013
Cone, J. D. (1999). Introduction to the special section on self-monitoring: A major assessment method in clinical psychology. *Psychological Assessment, 11,* 411–414.
Conners, C. K. (2008). *Conners' Third Edition.* North Tonawanda, NY: Multi-health Systems.
Conway, A.R.A.,Getz, S. J.,Macnamara, B.,&Engel De Abreu, P. M. J. (2011). Working memory and intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 394–418). New York: Cambridge University Press.
Corrigan, P. W., Hess, L., & Garman, A. N. (1998). Results of a job analysis of psychologists working in state hospitals. *Journal of Clinical Psychology, 54,* 11–18.
Cosden,M. (1984). Piers-Harris Children's Self-Concept Scale. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques, vol. I* (pp. 511–521). Kansas City, MO: Test Corporation of America.
Cosden, M. (1985). Rotter Incomplete Sentences Blank. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques, vol. II* (pp. 653–660). Kansas City, MO: Test Corporation of America.
Cosden,M. (1995). Review of Draw A Person: Screening Procedure for Emotional Disturbance. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 321–322). Lincoln: University of Nebraska Press.
Costa, P. T. Jr., & Mccrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources, Inc.
Cowan, N. (2005). Working memory capacity. New York: Psychology Press.
Craig, R. J., & Horowitz, M. (1990). Current utilization of psychological tests at diagnostic practicum sites. *The Clinical Psychologist, 43*(20), 29–36.
Crawford V. Honig, 37 F.3d 485 (9th Cir. 1994).
Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.
Cronbach, L. J. (1949). *Essentials of psychological testing.* New York: Harper & Brothers.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.
Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: HarperCollins.
Crosson, B., &Warren, R. L. (1982). Use of the Luria-Nebraska Neuropsychological Battery in aphasia: A conceptual critique. *Journal of Consulting and Clinical Psychology, 50,* 22–31.
Csikzentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience sampling method. *The Journal of Nervous and Mental Diseases, 175,* 526–533.
Culross, R. R., & Nelson, S. (1997). Training in personality assessment in specialist-level school psychology programs. *Psychological Reports, 81,* 119–124.
Cummings, J. A. (1995). Review of the Woodcock-Johnson Psycho-Educational Battery—Revised. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 1113–1116). Lincoln: Buros Institute of Mental Measurements.
Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1960). *An MMPI handbook: A clinical interpretation* (Vol. 1). Minneapolis: University of Minnesota Press.
Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook: Research applications* (Vol. 2). Minneapolis: University of Minnesota Press.
Dalessandro, S. P., Stilwell, L. A., Lawlor, J.A.,&Reese, L. M. (2010). *LSAT Performance with Regional, Gender, and Racial/Ethnic Breakdowns: 2003–2004 Through 2009–2010 Testing Years: LSAT Technical Report 10-03.* Newtown, PA: Law School Admission Council.
Dana, R. H. (1978). Review of the Rorschach. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 1040–1042). Lincoln: University of Nebraska Press.
Dana, R. H. (1996). The Thematic Apperception Test (TAT). In C. S. Newmark (Ed.), *Major psychological assessment instruments* (2nd ed., pp. 166–205). Boston: Allyn and Bacon.
Darwin, C. (1859). *On the origin of species by means of natural selection.* London: J. Murray.
Darwin, C. (1871). *The descent of man, and selection in relation to sex.* London: J. Murray.

Darwin, C. (1872). *The expression of the emotions in man and animals.* London: J. Murray.
Das, J. P. (1994). Serial and parallel processing. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 964–966). New York: Macmillan.
Das, J. P.,Naglieri, J. A., & Kirby, J.R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence.* Boston: Allyn and Bacon.
Data Research. (1997). *Students with disabilities and special education* (14th ed.). Rosemont, MN: Author.
Daw, J. (2001). Psychological assessments shown to be as valid asmedical tests. *Monitor on Psychology, 12*(7), 46–47.
Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: Free Press.
De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford.
Deary, I. J., & Batty, G. D. (2011) Intelligence as a predictor of health, illness, and death. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 683–707). New York: Cambridge University Press.
Deary, I. J., & Stough, C. (1996). Intelligence and inspection time. *American Psychologist, 51,* 599–608.
Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology, 86*(1), 130–147.
Debra P. ex rel. Irene P. V. Turlington, 730 F.2d 1405 (11th Cir. 1984).
Defense Manpower Data Center. (2004). *ASVAB Norms for the Career Exploration Program.* Washington, DC: Author.
Derogatis, L. R. (1994). *SCL-90-R Symptom Checklist- 90-R: Administration, scoring, and procedures manual* (3rd ed.). Minneapolis, MN: NCS Pearson.
Destefano, L. (2001). Review of the Otis-Lennon School Ability Test, Seventh Edition. In B. S. Plake & J. C. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 875–879). Lincoln: University of Nebraska Press.
Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95,* 542–575.
Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist, 55,* 34–43.
Digman, J.M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417–440.
Dikli, S. (2006).An overviewof automated scoring of essays. Journal of Technology, *Learning, and Assessment, 5*(1). Retrieved June 13, 2011, from http://www.jtla.org
Doll, E. A. (1935). A genetic scale of social maturity. *The American Journal of Orthopsychiatry, 5,* 180–188.
Doll, E. A. (1965). *Vineland Social Maturity Scale.* Circle Pines, MN: American Guidance Service.
Donnay, D. A. C. (1997). E. K. Strong's legacy and beyond: 70 years of the Strong Interest Inventory. *The Career Development Quarterly, 46,* 2–22.
Donnay, D. A. C., Morris, M. L., Schaubhut, N. A., & Thompson, R. C. (2005). *Strong Interest Inventory manual: Research, development, and strategies for interpretation.* Mountain View, CA: CPP, Inc.
Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores,* College Board Report No. 99-1. New York: College Entrance Examination Board.
Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACTAssessment and recentered SAT I sum scores. *College & University, 73*(2), 24–33.
Drasgow, F., Luecht, R. M., & Bennett, R. (2004). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–516). Westport, CT: Praeger.
Dreger, R.M. (1978). Review of the State-Trait Anxiety Inventory. In O. K. Buros (Ed.), *Eighth mental measurements yearbook* (pp. 1088–1095). Lincoln: University of Nebraska Press.
Dubois, P. H. (1970). *A history of psychological testing.* Boston, MA: Allyn & Bacon, Inc.
Duckworth, J. C., & Levitt, E. E. (1994). Minnesota Multiphasic Personality Inventory-2. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques,* vol. X (pp. 424–428). Kansas City, MO: Test Corporation of America.
Dunn, L. M., & Dunn, D. M. (1981). *Peabody Picture Vocabulary Test—Revised Manual.* Circle Pines, MN: American Guidance Service.
Dunn, L. M., & Dunn, D. M. (1997a). *Peabody Picture Vocabulary Test, Third Edition, Examiner's Manual.*

Circle Pines, MN: American Guidance Service.
Dunn, L. M., & Dunn, D. M. (1997b). *Peabody Picture Vocabulary Test*, *Third Edition, Norms Booklet.* Circle Pines, MN: American Guidance Service.
Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test, Fourth Edition, Examiner's Manual.* San Antonio, TX: Pearson.
Dupaul,G. J., Power, T. J.,Anastopoulos, A.D.,&Reid, R. (1998). *ADHD Rating Scale IV.* Los Angeles: Western Psychological Services.
Durlak, J. A. (1996). Understanding meta-analysis. In L. G. GRIMM & P. R. YARNOLD (Eds.), *Reading and understanding multivariate statistics* (pp. 319–352). Washington, DC: American Psychological Association.
Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
Editorial Board. (1996). Definition of mental retardation. In J. W. Jacobson & J. A. Mulick (Eds.), *Manual of diagnosis and professional practice in mental retardation* (pp. 13–53). Washington, DC: American Psychological Association.
Educational Testing Service. (2000). *Major field tests: Comparative data guide and descriptions of reports.* Princeton, NJ: Author.
Educational Testing Service. (2012). About ETS 1 Major Field Tests. Retrieved June 15, 2012, from http:// www.ets.org/mft/about
Edwards, A. L. (1957). *Techniques for attitude scale construction.* New York: Appleton-Century-Crofts.
Edwards, A. L. (1959). *Edwards Personal Preference Schedule Manual,* Revised. New York: The Psychological Corporation.
Eid, M., & Larsen,R. J. (2008). *The science of subjective wellbeing.* New York: Guilford.
Engelhard, G., Davis, M., & Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education, 12*(4), 199–210.
Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3*(4), 347–360.
Engler, B. (1999). *Personality theories: An introduction* (5th ed.). Boston: Houghton Mifflin.
Epstein, J. H. (1985). Review of the Piers-Harris Children's Self-Concept Scale (TheWay I Feel About Myself). In J.V. Mitchell, Jr. (Ed.), *The ninth mental measurements yearbook* (pp. 1167–1169). Lincoln: University of Nebraska Press.
Equal Educational Opportunity Commission. (1978). *Uniform guidelines on employee selection—29CFR1607.18.* Washington, DC: U.S. Government Printing Office.
Erdberg, P. (1996). The Rorschach. In C. S. Newmark (Ed.), *Major psychological assessment instruments* (2nd ed., pp. 148–l65). Boston: Allyn and Bacon.
Esquivel, G. B. (1984). Coloured Progressive Matrices. In D. Keyser & R. Sweetland (Eds.), *Test critiques,* Vol. I (pp. 206–213). Austin, TX: PRO-ED.
ETS. (2005). *Graduate Record Examinations 2005–2006 guide to the use of scores.* Princeton, NJ: Author.
ETS. (2007). *The GRE Analytical Writing Measure: An asset in admissions decisions.* Princeton, NJ: Author.
ETS. (2009). *A Comprehensive Review of Published GRE Validity Data: A summary from ETS.* Princeton, NJ: Author.
ETS. (2012a). *Graduate Record Examinations guide to the use of scores 2012–2013.* Princeton, NJ: Author.
ETS.(2012b). *GRE Bulletin2012–13.* Princeton, NJ: Author.
Ewing, M., Huff, K., Andrews, M., & King, K. (2005). *Assessing the reliability of skills measured by the SAT* (RN-24). New York: College Board.
Exner, J. E. (1991). The *Rorschach: A comprehensive system. Volume 2: Interpretation* (2nd ed.). New York: Wiley.
Exner, J. E. (1993). *The Rorschach: A comprehensive system. Volume 1: Basic foundations* (3rd ed.). New York: Wiley.
Exner, J. E., & Weiner, I. B. (1995). *The Rorschach: A comprehensive system. Volume 3: Assessment of children and adolescents* (2nd ed.). New York: Wiley.
Exner, J. E., Jr. (2003). *The Rorschach: A comprehensive system. Vol. 1: Basic foundations and principles of interpretation* (4th ed.). New York: Wiley.
Exner, J. E., Jr., & Erdberg, P. (2005). *The Rorschach: A comprehensive system. Vol. 2: Advanced*

interpretation (3rd ed.). New York: Wiley.
Eyde, L. D.,Moreland, K. L.,Robertson, G. J., Primoff, E. S., & Most, R. B. (1988). *Test user qualifications: A data-based approach to promoting good test use.* Washington, DC: American Psychological Association.
Eysenck, H. J. (1970). The structure of human personality. London: Methuen. FARMER, R. F. (2001). Review of the Beck Depression Inventory-II. In B. S. Plake & J. C. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 123–126). Lincoln: University of Nebraska Press.
Fear, R. A. (2002). *The evaluation interview: How to probe deeply, get candid answers, and predict the performance of job candidates.* New York: McGraw-Hill.
Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education/Oryx.
Ferrara, S. (1998). Review of the Wechsler Individual Achievement Test. In J. C. Impara & B. S. Plake (Eds.), *The thirteenth mental measurements yearbook* (pp. 1128–1132). Lincoln: University of Nebraska Press.
Finger, M. S., & Ones, D. S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment, 11*(1), 58–66.
Finn, S. E. (1996). *Manual for using the MMPI-2 as a therapeutic intervention.* Minneapolis: University of Minnesota Press.
First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *User's guide for the Structured Clinical Interview for DSM-IV Axis I Disorders, Clinical Version.* Washington, DC: American Psychiatric Press.
Fischer, J., & Corcoran, K. (2000). *Measures for clinical practice: A sourcebook* (3rd ed., Vols. 1–2). New York: Free Press.
Fischer, J., & Corcoran, K. (2007). *Measures for clinical practice and research: A sourcebook* (4th ed., Vols. 1–2). New York: Oxford.
Fisher, C. B. (2010). *Decoding the ethics code: A practical guide for psychologists, updated* (2nd ed.). Thousand Oaks, CA: Sage.
Fiske, D. W., & Campbell, D. T. (1992). *Citations do not solve problems. Psychological Bulletin, 112,* 393–395.
Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.
Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171– 191.
Flynn, J. R. (1994). IQ gains over time. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 617–623). New York: Macmillan.
Flynn, J.R. (1999). Searching for justice:The discoveryof IQ gains over time. *American Psychologist, 54,* 5–20.
Flynn, J. R. (2011). Secular changes in intelligence. In J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 647–665). New York: Cambridge University Press.
Folstein,M. F., Folstein, S. E., &Mchugh, P.R. (1975).Mini-mental state. *Journal of Psychiatric Research, 12,* 189–198.
Folstein, M. F., Folstein, S. E., Mchugh, P. R., & Fanjiang, G. (2000). *Mini-Mental State Examination user's guide.* Odessa, FL: Psychological Assessment Resources.
Folstein,M. F., Folstein, S. E.,White, T.,&Messer, M. (2010). *Mini-Mental State Examination* (2nd ed.) user's manual. Lutz, FL: PAR.
Foxhall, K. (2000). Bringing law and psychology together. *Monitor on Psychology, 31*(1).
Franzen, M. D. (2000). *Reliability and validity in neuropsychological assessment.* New York: Plenum.
Frauenhoffer, D., Ross, M. J., Gfeller, J., Searight, H. R., & Piotrowski, C. (1998). Psychological test usage among licensed mental health practitioners: A multidisciplinary survey. *Journal of Psychological Practice, 4,* 28–33.
Gadow, K. D., & Sprafkin, J. (1997). *Attention Deficit Hyperactivity Disorder Symptom Checklist-4* (ADHD–SC4). Stony Brook, NY: Checkmate Plus.
Gallagher, A. M., & Kaufman, J. C. (2005). *Gender differences in mathematics.* New York: Cambridge University Press.
Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences.* London: Macmillan.
Galton, F. (1883). *Inquiries into human faculty and its development.* London: Macmillan.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.
Gardner, H. (1986). The waning of intelligence tests. In R. J. Sternberg & D.K. Detterman (Eds.), *What is intelligence?* (pp. 73–76). Norwood, NJ: Ablex Publishing.
Gardner, H. (1993). *Multiple intelligences: The theory in practice.* New York: Basic Books.
Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century.* New York: Basic Books.
Gardner, H. (2006). *Five minds for the future.* Boston: Harvard Business School Publishing.
Garner, D. M. (1991). *Eating Disorder Inventory-2 professional manual.* Odessa, FL: Psychological Assessment Resources.
Garner,D.M. (2004). *Eating Disorder Inventory-3 professional manual.* Lutz, FL: Psychological Assessment Resources.
Gi Forum Images de Tejas v. Texas Educ. Agency, 87 F. Supp.2d 667 (W.D. Tex 2000).
Gieser, L., & Stein, M. I. (Eds.) (1999). *Evocative images: The Thematic Apperception Test and the art of projection.* Washington, DC: American Psychological Association.
Glass, C. R., & Arnkoff, D. B. (1997). Questionnaire methods of cognitive self-statement assessment. *Journal of Consulting and Clinical Psychology, 65,* 911–927.
Glass, G.V.,&Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn and Bacon.
Golden, C. (1978). *Stroop Color and Word Test: Manual for clinical and experimental uses.* Chicago: Stoelting.
Golden, C. J. (1981). A standardized version of Luria's neuropsychological tests. In S. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology.* New York: Wiley-Interscience.
Golden, C. J. (1984). Applications of the standardized Luria-Nebraska Neuropsychological Battery to rehabilitation planning. In P. E. Logue & J. M. Schear (Eds.), *Clinical neuropsychology: A multidisciplinary approach.* Springfield, IL: C. C. Thomas.
Golden, C. J.,Hammeke, T. A., & Purisch, A. D. (1979). Diagnostic validity of a standardized neuropsychological battery derived from Luria's neuropsychological tests. *Journal of Consulting and Clinical Psychology, 46,* 1258–1265.
Golden, C. J., Purisch, A.D., & Hammeke, T. A. (1985). *Manual for the Luria-Nebraska Neuropsychological Battery: Forms I and II.* Los Angeles, CA: Western Psychological Services.
Goldfried, M. R. (1976). Behavioral assessment. In I. B. Weiner (Ed.), *Clinical methods in psychology* (pp. 281–330). New York: Wiley.
Goldfried, M. R., & Kent, R. N. (1972). Traditional vs. behavioral assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin, 77,* 409–420.
Goldman, B. A. (2001). Review of the Otis-Lennon School Ability Test, Seventh Edition. In B. S. Plake & J. C. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 879–881). Lincoln, NE: University of Nebraska Press.
Goldman, B. A., & Mitchell, D. F. (2003). *Directory of unpublished experimental mental measures* (Vol. 8). Washington, DC: American Psychological Association.
Goldman, B. A., & Mitchell, D. F. (2008). *Directory of unpublished experimental mental measures* (Vol. 9). Washington, DC: American Psychological Association.
Goode, D. (2002). Mental retardation is dead: Long live mental retardation. *Mental Retardation, 40*(1), 57–59.
Goodenough, F. (1926). *Measurement of intelligence by drawings.* New York: World Book.
Gottfredson, L. S. (1996). What do we know about intelligence? *American Scholar, Winter,* 15–30.
Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13–23.
Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31* (4), 343–398.
Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology, 86*(1), 174–199.
Graham, J. R. (2006). *MMPI-2: Assessing personality and psychopathology* (4th ed.). New York: Oxford University Press.
Graham, J. R. (2011). *MMPI-2: Assessing personality and psychopathology* (5th ed.). New York: Oxford University Press.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17*(2), 16–19.
Green, D. R., Yen, W. M., & Burkett, G. R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education, 2,* 297–312.
Greene, R. L. (2010). *The MMPI-2/MMPI-2-RF: An interpretive manual* (3rd ed.). San Antonio, TX: Pearson.
Gregory, R. J. (1996). *Psychological testing: History, principles, and applications* (2nd ed.). Boston: Allyn and Bacon.
Gregory, R. J. (2011). *Psychological testing: History, principles, and applications* (6th ed.). Boston: Pearson.
Griggs v. Duke Power Co., 401 U.S. 424 (1971).
Grisso, T. (1986). Psychological assessment in legal contexts. In W. J. Curran, A. L. McGarry, & S. A. Shah (Eds.), *Forensic psychiatry and psychology.* Philadelphia, PA: F. A. Davis.
Grisso, T. (1996). Clinical assessments for legal decisionmaking in criminal cases: Research recommendations. In B. D. Sales & S. A. Shah (Eds.), *Mental health and law: Research, policy, and services* (pp. 109–140). Durham, NC: Carolina Academic Press.
Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research* (2nd ed.). New York: Routledge.
Gross, M. (1962). *The brain watchers.* New York: Random House.
Groth-Marnat, G. (1999). *Handbook of psychological assessment* (3rd ed.). New York: Wiley.
Groth-Marnat, G. (2003). *Handbook of psychological assessment* (4th ed.). New York: Wiley.
Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). New York: Wiley.
Group for the Advancement of Psychiatry. (1991). *The mental health professional and the legal system.* New York: Bruner/Mazel.
Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.
Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.
Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.
Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin, 53,* 267–293.
Guilford, J. P. (1958). A system of psychomotor abilities. *American Journal of Psychology, 71,* 164–174.
Guilford, J. P. (1959a). *Personality.* New York: McGraw Hill.
Guilford, J. P. (1959b). Three faces of intellect. *American Psychologist, 14,* 469–479.
Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.
Guilford, J. P. (1985). The structure of intellect model. In B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 225–266). New York: Wiley.
Guilford, J. P. (1988). Some changes in the structure-ofintellect model. *Educational and Psychological Measurement, 48,* 1–4.
Gulliksen, H. (1950). *Theories of mental tests.* New York: Wiley.
Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150.
Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement, 7,* 247–280.
Guttman, L., & Suchman, E. A. (1947). Intensity and a zero point for attitude analysis. *American Sociological Review, 12,* 57–67.
Haladyna, T.M. (1994). *Developing and validating multiple-choice items.* Hillsdale, NJ: Lawrence Erlbaum.
Haladyna, T.M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Erlbaum.
Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Erlbaum.
Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2,* 37–50.
Haladyna, T. M.,& Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2,* 51–78.
Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309–334.
Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Erlbaum.

Halpern, D. F., Beniner, A. S.,&Straight, C. A. (2011). Sex differences in intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 253–272). New York: Cambridge University Press.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89– 116). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, NJ: Sage Publications.

Handler, L. (1996). The clinical use of figure drawings. In C. S. Newmark (Ed.), *Major psychological assessment instruments* (2nd ed., pp. 206–293). Boston: Allyn and Bacon.

Harcourt Educational Measurement. (2003). *Stanford Achievement Test Series, Tenth Edition: Spring technical data report.* San Antonio, TX: Author.

Harmon, L. W., Hansen, J. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong Interest Inventory: Applications and technical guide.* Palo Alto, CA: Consulting Psychologists Press.

Harris, D.B. (1963). *Children's drawings as measures of intellectual maturity.* New York: Harcourt, Brace, & World.

Harrow, A. J. (1972). *A taxonomy of the psychomotor domain.* New York: David McKay.

Harwood, T. M., Beutler, L. E., & Groth-Marnot, G. (2011). *Integrative assessment of adult personality* (3rd ed.). New York: Guilford.

Hathaway, S. R., & Mckinley, J. C. (1989). *MMPI-2 manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Hayes, J. R., Hatch, J. A., & Silk, C. M. (2000). Does holistic assessment predict writing performance? *Written Communication, 17*(1), 3–26.

Haynes, S. N. (2001). Introduction to the special section on clinical applications of analogue behavioral observation. *Psychological Assessment, 13,* 3–4.

Haynes, S. N., & Kaholokula, J. K. (2008). Behavioral assessment. In M. Hersen & A. M. Gross (Eds.), *Handbook of clinical psychology* (Vol. 1: Adults, pp. 495–522). Hoboken, NJ: Wiley.

Haynes, S. N.,O'brien,W. H.,&Kaholokula, J. K. (2011). *Behavioral assessment and case formulation.* Hoboken, NJ: Wiley.

Haynes, S. N., Smith, G. T., & Hunsley, J. D. (2011). *Scientific foundations of clinical assessment.* New York: Routledge.

Hayslip, B. (1994). Stability of intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 1019–1026). New York: Macmillan.

Helliwell, J., Layard, R., & Sachs, J. (2012). *World happiness report.* Retrieved from http://issuu.com/earthinstitute/docs/world-happiness-report

Henryssen, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130–l59). Washington, DC: American Council on Education.

Herk, N. A., & Thompson, R. C. (2012). *Strong Interest Inventory manual supplement.* Mountain View, CA: CPP, Inc.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: The Free Press.

Hertz, M. A. (1943). Personality patterns in adolescence as portrayed by the Rorschach ink blot method: IV The "Erlebnistypus." *Journal of General Psychology, 20,* 3–45.

Hertz, M. A. (1948). Suicidal configurations in Rorschach records. *Rorschach Research Exchange, 12,* 3–58.

Hertzog, C. (2011). Intelligence in adulthood. In J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 174–190). New York: Cambridge University Press.

Hess, A. K. (1998). Review of the Millon Clinical Multiaxial Inventory—III. In B. S. Plake & J. C. Impara (Eds.), *The thirteenth mental measurements yearbook* (pp. 665–667). Lincoln: University of Nebraska Press.

Hilgard, E. R. (1987). *Psychology in America: A historical survey.* San Diego, CA: Harcourt Brace Jovanovich.

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-

analysis of Rorschach and MMPI validity. *Psychological Assessment, 11*(3), 278–296.
Hilsenroth, M. J., & Handler, L. (1995). A survey of graduate students' experiences, interests, and attitudes about learning the Rorschach. *Journal of Personality Assessment, 64,* 243–257.
Hoffman, B. (1962). *The tyranny of testing.* New York: Crowell-Collier.
Hogan, T. P. (1981). *Relationship between free-response and choice-type tests of achievement: A review of the literature.* Green Bay: University of Wisconsin. (ERIC Document Reproduction Service No. ED 224 811.).
Hogan, T. P. (2005a). A list of 50 widely used psychological tests. In G. P. Koocher, J. C. Norcross, & S. S. Hill (Eds.), *Psychologist's Desk Reference* (2nd ed., pp. 101–104). New York: Oxford University Press.
Hogan, T. P. (2005b). Types of test scores and their percentile equivalents. In G. P. Koocher, J. C. Norcross, & S. S. Hill (Eds.), *Psychologists' desk reference* (2nd ed.). New York: Oxford University Press.
Hogan, T. P. (2007). *Educational assessment: A practical introduction.* New York: Wiley.
Hogan, T. P. (2013). Constructed-response approaches for classroom assessment. In J. R. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 275–292). Thousand Oaks, CA: Sage.
Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64,* 802–812.
Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education, 20,* 427–441.
Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60,* 523–531.
Holaday, M., Smith, D. A., & Sherry, A. (2000). Sentence completion tests: A review of the literature and results of a survey of members of the Society for Personality Assessment. *Journal of Personality Assessment, 74,* 371–383.
Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6,* 35–45.
Holland, J. L. (1966). *The psychology of vocational choice: A theory of personality types and model environments.* Waltham, MA: Ginn.
Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.
Holland, J. L., Fritzsche, B. A., & Powell, A. B. (1997). *SDS Self-directed Search Technical manual.* Odessa, FL: Psychological Assessment Resources.
Holland, J. L., Powell, A. B., & Fritzsche, B. A. (1997). *SDS Self-directed Search professional user's guide.* Odessa, FL: Psychological Assessment Resources.
Holland, P. W.,&Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H.Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129– 145). Hillsdale, NJ: Lawrence Erlbaum.
Holtzman, W. H. (1961). *Holtzman Inkblot Technique administration and scoring guide.* New York: Psychological Corporation.
Horn, J. L. (1994). Fluid and crystallized intelligence, theory of. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 443–451). New York: Macmillan.
Horn, J. L., & CATTELL, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57,* 253–270.
House, A. E. (1996). The Wechsler Adult Intelligence Scale—Revised (WAIS-R). In C. S. Newmark (Ed.), *Major psychological assessment instruments* (2nd ed., pp. 320–347). Needham Heights, MA: Allyn & Bacon.
Hoyer, W. J., & Touron, D. R. (2003). Learning in adulthood. In J. Demick & C. Andreoletti (Eds.), *Handbook of adult development* (pp. 23–41). New York: Kluwer Academic/Plenum.
Huffcutt, A. I. (2011). An empirical review of the employment interview construct literature. *International Journal of Selection and Assessment, 19*(1), 62–81.
Huffcutt, A. I., & Youngcourt, S. S. (2007). Employment interviews. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 181–199). New York: Taylor and Francis.
Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment, 11*(3), 266–277.
Hunsley, J., & Haynes, S. N. (Eds.). (2003). Special section: Incremental validity and utility in clinical assessment. *Psychological Assessment, 15,* 443–531.

Hunsley, J., & Mash, E. J. (Eds.). (2008). *A guide to assessments that work*. New York: Oxford.
Hunt, E. (2011). *Human intelligence.* New York: Cambridge University Press.
Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721–735.
Hutton, J. B.,Dubes, R., & Muir, S. (1992). Assessment practices of school psychologists: Ten years later. *School Psychology Review, 21,* 271–284.
Impara, J. C. (Ed.). (1995). *Licensure testing: Purposes, procedures, and practices.* Lincoln, NE: Buros Institute of Mental Measurements.
Jencks, C. (1979). Who gets ahead? *The determinants of economic success in America.* New York: Basic Books.
Jencks, C., & Phillips, M. (Eds.). (1998). *The black-white test score gap.* Washington, DC: Brookings Institution Press.
Jenkins, S. R. (Ed.). (2008). *A handbook of clinical scoring systems for thematic apperceptive techniques.* New York: Lawrence Erlbaum Associates.
Jensen, A. R. (1969). *How much can we boost IQ and scholastic achievement?* Harvard Educational Review, 39, 1–123.
Jensen, A.R. (1980). Bias in mental testing. New York: Free Press.
Jensen, A. R. (1994). Race and IQ scores. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 899–907). New York: Macmillan.
Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.
Jeske, P. J. (1985).Reviewof the Piers-HarrisChildren's Self- Concept Scale (The Way I Feel About Myself). In J. V. Mitchell, Jr. (Ed.), *The ninth mental measurements yearbook* (pp. 1169–1170). Lincoln: University of Nebraska Press.
Johnson, S. T. (1994). Scholastic Assessment Tests (SAT). In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 956–960). New York. Macmillan.
Johnson, W. L., & Dixon, P. N. (1984). Response alternatives in Likert scaling. Educational and *Psychological Measurement, 44,* 563–567.
Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: Author.
Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education—revised.* Washington, DC: Author.
Jones, J. L., & Mehr, S. L. (2007). Foundations and assumptions of the scientist-practitioner model. *American Behavioral Scientist, 50*(6), 766–771.
Journal of Educational, Psychology. (1921). Intelligence and its measurement: A symposium. Editorial introduction. *Journal of Educational Psychology, 12,* 123.
Juni, S. (1995). Review of the Revised NEO Personality Inventory. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 863–868). Lincoln: University of Nebraska Press.
Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342.
Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications, and issues* (6th ed.). Belmont, CA: Wadsworth/Thomson Learning.
Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological testing: Principles, applications, and issues* (8th ed.). Belmont, CA: Wadsworth/Thomson Learning.
Karraker v. Rent-A-Center, Inc., 411 F.3d 831 (7th Cir. 2005).
Katin, E. S. (1978). Review of the State-Trait Anxiety Inventory. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 1095–1096). Lincoln: University of Nebraska Press.
Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children* (2nd Ed.). Circle Pines, MN: American Guidance Service.
Kelley, M. L. (2005). Review of Piers-Harris Children's Self-Concept Scale. In R. A. Spies, B. S. Plake, & L. L. Murphy (Eds.), *The sixteenth mental measurements yearbook* (pp. 789–790). Lincoln, NE: Buros Institute of Mental Measurements.
Kelley, T. L. (1927). *Interpretation of educational measurements.* Yonkers-on-Hudson, NY: World Book.
Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17–24.

Kennedy, M. L., Faust, D., Willis, W. G., & Piotrowski, C. (1994). Socio-emotional assessment practices in school psychology. *Journal of Psychoeducational Assessment, 12,* 228–240.

Keyes, C. L. M.,&Magyar-Moe, J. L. (2003). Measurement and utility of adult subjective well-being. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 411–425). Washington, DC: American Psychological Association.

Keyser, D. J. (1994). *Test critiques* (Vol. XI). Austin, TX: PRO-ED.

Keyser, D. J. (Ed.). (2004). *Test Critiques* (Vol. XI). Austin, TX: PRO-ED.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107,* 296–310.

Kline, P. (1991). *Intelligence: The psychometric view.* New York: Routledge.

Kline, P. (1994). Cattell, R. B. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 241–243). New York: Macmillan.

Klopfer, B. (1937). The present status of the theoretical development of the Rorschach method. *Rorschach Research Exchange, 1,* 142–147.

Klopfer, B., & Kelley, D. (1942). *The Rorschach technique.* Yonkers, NY: World Book.

Klopfer, W. G., & Taulbee, E. S. (1976). *Projective tests. Annual Review of Psychology, 27,* 543–568.

Knabb, J. J., Vogt, R. G., & Newgren, K. P. (2011). MMPI-2 characteristics of the Old Order Amish: A comparison of clinical, nonclinical, and United States normative samples. *Psychological Assessment, 23,* 865–875.

Knapp, J. E., & Knapp, L. G. (1995). Practice analysis: Building the foundation for validity. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 93–116). Lincoln: University of Nebraska.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S.M. (2008). *Validity of the SAT for predicting first-year college grade point average, College Board Research Report No. 2008-5.* New York: College Board.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Koocher, G. P., & Keith-Spiegel, P. (1998). *Ethics in psychology: Professional standards and cases* (2nd ed.). New York: Oxford University Press.

Koocher, G. P., & Keith-Spiegel, P. (2008). *Ethics in psychology and the mental health professions: Standards and cases* (3rd ed.). New York: Oxford.

Korotitsch, W. J., & Nelson-Gray, R. O. (1999). An overview of self-monitoring research in assessment and treatment. *Psychological Assessment, 11,* 415–425.

Krathwohl, D. R., Bloom, B. S.,&Masia, B. B. (1964). *Taxonomy of educational objectives, handbook II: Affective domain.* New York: David McKay.

Kreitzer, A. E., & Madaus, G. F. (1994). Empirical investigations of the hierarchical structure of the taxonomy. In L. W. Anderson & L. A. Sosniak (Eds.), *Bloom's taxonomy: A forty-year retrospective* (pp. 64–81). Chicago: University of Chicago Press.

Kuder, F., & Zytowski, D. G. (1991). *Kuder Occupational Interest Survey, Form DD, General Manual* (3rd ed.). Adel, IA: National Career Assessment Services.

Kuder, J. F., & Richardson, M.W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151–160.

Kuncel, N. R.,&Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19,* 339–345.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*(1), 148–161.

Kuncel, N. R.,Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the graduate record examination for master's and doctoral programs:Ameta-analytic investigation. *Educational and Psychological Measurement, 70,* 340–352.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 24–28.
Larry P. ex rel. Lucille P. v. Riles, 793 F.2d 969 (9th Cir. 1984).
Larson, G. E. (1994). Armed Services Vocational Aptitude Battery. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 121–l24). New York: Macmillan.
Lawrence, I., Rigol, G. W., Van Essen, T., & Jackson, C. A. (2002). *A historical perspective on the SAT 1926–2001.* New York: College Board.
Lawshe, C. H. (1978). A quantitative approach to content validity. *Personnel Psychology, 28,* 563–575.
Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31,* 3–12.
Lee, S.W.,& Stefany, E. F. (1995). Reviewof the Woodcock-Johnson Psych-Educational Battery—Revised. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 1116–1117). Lincoln: Buros Institute of Mental Measurements.
Lee, Y. W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing, 7,* 353–385.
Lees-Haley, P. R. (1992). Psychodiagnostic test usage by forensic psychologists. *American Journal of Forensic Psychology, 10*(1), 25–30.
Lees-Haley, P. R., English, L., & Glenn, W. (1991). A fake bad scale on the MMPI-2 for personal injury claimants. *Psychological Reports, 68,* 203–210.
Lees-Haley, P. R., Smith, H. H., Williams, C. W., & Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical study. *Archives of Clinical Neuropsychology, 11,* 45–51.
Lennon, R. T. (1985). Group tests of intelligence. In B. B. Wolman (Ed.), Handbook of intelligence: *Theories, measurements, and applications* (pp. 825–845). New York: Wiley.
Lerner, H., & Lerner, P. M. (1985). Rorschach Inkblot Test. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques,* vol. IV (pp. 523–552). Kansas City, MO: Test Corporation of America.
Levitt, E. E., & Gotts, E. E. (1995). *The clinical application of MMPI special scales* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Lewak, R. W., Marks, P. A., & Nelson, G. E. (1990). *Therapist's guide to the MMPI and MMPI-2: Providing feedback and treatment.* Muncie, IN: Accelerated Development.
Lezak, M. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.
Licht, M. H. (1995). Multiple regression and correlation. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 19–64). Washington, DC: American Psychological Association.
Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 1–55.
Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.
Lindenberger, U., & Baltes, P. B. (1994). Aging and intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 52–66). New York: Macmillan.
Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16.
Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 28–30.
Linn, R. L. (2000). Assessment and accountability. *Educational Researcher, 29*(2), 4–16.
Linn, R. L., & Miller, D. (2004). *Measurement and assessment in teaching* (9th ed.). Englewood Cliffs, NJ: Prentice-Hall.
Llabre, M. M. (1984). Standard Progressive Matrices. In D. Keyser & R. Sweetland (Eds.), *Test critiques*, Vol. I (pp. 595–602). Austin, TX: PRO-ED.
Loehlin, J. C. (1994). Genetics, behavior. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 475– 483). New York: Macmillan.
Lopez, S. J., & Snyder, C. R. (Eds.). (2003). *Positive psychological assessment: A handbook of models and measures.* Washington, DC: American Psychological Association.
Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Lubin, B., Larsen, R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States:

1935–1982. *American Psychologist, 39,* 451–453.
Lubin, B., Larsen, R. M., Matarazzo, J. D., & Seever, M. (1985). Psychological test usage patterns in five professional settings. *American Psychologist, 49,* 857–861.
Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's" 'General intelligence,' objectively determined and measured." *Journal of Personality and Social Psychology, 86*(1), 96–111.
Macan, T. (2009). The employment interview: A review of current studies and directions for future research. *Human Resource Management Review, 19*(3), 203–218.
Machover, K. (1949). *Personality projection in the drawing of the human figure.* Springfield, IL: Charles C Thomas.
Mackintosh, N. J. (1998). *IQ and human intelligence.* New York: Oxford University Press.
Macmillan, M. (2000). *An odd kind of fame: Stories of Phineas Gage.* Cambridge, MA: MIT Press.
Maddox, T. (2003). *Tests: A comprehensive reference for assessments in psychology, education, and business* (5th ed.). Austin, TX: PRO-ED.
Marlowe, D. B.,Wetzler, S., & Gibbings, E. N. (1992). Graduate training in psychological assessment: What Psy.D. 's and Ph.D. 's must know. *The Journal of Training and Practice in Professional Psychology, 6,* 9–18.
Marsh, H. W., Parada, R. H., & Ayotte, V. (2004). A multidimensional perspective of relations between self-concept (Self Description Questionnaire II) and adolescent mental health (Youth Self Report). *Psychological Assessment, 16,* 27–41.
Masterstrack.com. (2012). Ida Keeling at 97 becomes oldest American female sprinter. Retrieved June 12, 2012, from http://masterstrack.com/2012/06/22606/#more-22606
Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology, 56,* 506–509.
Mather, N., & Woodcock, R. W. (2001). *Woodcock- Johnson III Tests of Achievement: Examiner's manual.* Itasca, IL: Riverside Publishing.
Maxey, J. (1994). American College Test. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 82–85). New York: Macmillan.
Mccall, W. T. (1922). *How to measure in education.* New York: Macmillan.
Mcclelland, D. C. (1985). *Human motivation.* Glenview, IL: Scott, Foresman and Company.
Mcclelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive.* New York: Appleton Century Crofts.
Mccrae, R. R., Costa, P. T., Jr., & Martin, T. A. (2005). The NEO–PI–3: A more readable Revised NEO Personality Inventory. *Journal of Personality Assessment, 84*(3), 261–270.
Mccrae,R.R.,Martin, T. A.,&Costa, P. T., Jr. (2005). Age trends and age norms for the NEO Personality Inventory-3 in adolescents and adults. *Assessment, 12,* 363–373.
Mccrae, R. R., & Costa, P. T. (2010). *NEO Inventories for the NEO Personality Inventory-3 (NEO PI-3), NEO Five Factor Inventory-3 (NEO-FFI-3), NEO Personality Inventory-Revised (NEO PI-R) professional manual.* Lutz, FL: PAR.
Mckay, D. (Ed.). (2008). *Handbook of research methods in abnormal and clinical psychology.* Thousand Oaks, CA: Sage.
Mckhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group, Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34,* 939–944.
Mclellan, M. J. (1995). Review of the Rotter Incomplete Sentences Blank, Second Edition. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 882–883). Lincoln: University of Nebraska Press.
Mcmillan, J. H. (Ed.). (2013). *Sage handbook of research on classroom assessment.* Thousand Oaks, CA: Sage.
Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.
Mehrens,W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–18.

Messick, S. (1993). Validity. In R. L. LINN (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Phoenix, AZ: The Oryx Press.
Meyer, G. (Ed.). (1999). The utility of the Rorschach in clinical assessment [Special section: I]. *Psychological Assessment, 11,* 235–302.
Meyer, G. (Ed.). (2001). The utility of the Rorschach in clinical assessment [Special section: II]. *Psychological Assessment, 13,* 419–502.
Meyer, G. J. (Ed.). (2006). The MMPI-2 Restructured Clinical Scales [Special issue]. *Journal of Personality Assessment, 87*(2).
Meyer, G. J.,Viglione,D. J.,Mihura, J. L., Erard,R. E., & Erdberg, P. (2011). *Rorschach Performance Assessment System: Administration, coding, interpretation, and technical manual.* Toledo, OH: Rorschach Performance Assessment System.
Meyer, G. J., Finn, S. E., Eyde, L. D., et al. (2001). Psychological testing and psychological assessment. *American Psychologist, 56,* 128–165.
Mihura, J. L.,Meyer, G. J., Dumitrascu, N.,&Bombel, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the Comprehensive System. *Psychological Bulletin, 139,* 548–605.
Miller, M. D. (2010). Review of theWechsler Individual Achievement Test—Third Edition. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The eighteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.
Millman, J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). Phoenix, AZ: Oryx Press.
Millon, T. (1969). *Modern psychopathology: A biosocial approach to maladaptive learning and functioning.* Philadelphia: W. B. Saunders.
Millon, T. (1981). *Disorders of personality.* New York: Wiley.
Millon, T. (1994). *Millon Clinical Multiaxial Inventory—III manual.* Minneapolis, MN: National Computer Systems.
Millon, T. (2008). *The Millon inventories: A practitioner's guide to personalized clinical assessment* (2nd ed.). New York: Guilford Press.
Millon, T. (Ed.). (1997). *The Millon inventories: Clinical and personality assessment.* New York: Guilford Press.
Millon, T., & DAVIS, R. D. (1996). The Millon Clinical Multiaxial Inventory—III (MCMI-III). In C. S. New-Mark (Ed.), *Major psychological assessment instruments* (pp. 108–147). Needham Heights, MA: Allyn & Bacon.
Mischel, W. (1968). *Personality and assessment.* New York: Wiley.
Misiak, H. (1961). *The philosophical roots of scientific psychology.* New York: Fordham University Press.
Moosman, D. (2003). Atkins v. Virginia: A psychiatric can of worms. *New Mexico Law Review, 33,* 255–291.
Moreland, K. L., Eyde, L. D.,Robertson, G. J., Primoff, E. S., & Most, R. B. (1995). Assessment of test user qualifications: A research-based measurement procedure. *American Psychologist, 50*(1), 14–23.
Moreno, K. E., & Segall, D. O. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169–l74). Washington, DC: American Psychological Association.
Morrison, G. M. (1995). Review of Draw A Person: Screening Procedure for Emotional Disturbance. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 322–323). Lincoln: University of Nebraska Press.
Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17* (2), 6–12.
Murphy, G. (1949). *Historical introduction to modern psychology.* New York: Harcourt, Brace & World, Inc.
Murphy, K. (1984). Armed Services Vocational Aptitude Battery. In D. Keyser & R. Sweetland (Eds.), *Test critiques*, Vol. I (pp. 61–69). Austin, TX: PRO-ED.
Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (5th ed.).

Upper Saddle River, NJ: Prentice-Hall.
Murphy, L. L., Geisinger, K. F., Carlson, J. F., & Spies, R. A. (Eds.). (2011). *Tests in print VIII*. Lincoln, NE: Buros Center for Testing.
Murphy, L. L., Plake, B. S., Impara, J. C., & Spies, R. A. (2002). *Tests in print VI*. Lincoln, NE: University of Nebraska Press.
Murray, H. A. (1943). *Thematic Apperception Test.* Cambridge, MA: Harvard University Press.
Murray, H. A. et al. (1938). *Explorations in personality.* New York: Oxford University Press.
Myers, D. G. (2000). The funds, friends, and faith of happy people. *American Psychologist, 55,* 56–67.
Naglieri, J. A., & Das, J. P. (1997). *Cognitive Assessment System.* Itasca, IL: Riverside Publishing.
Naglieri, J. A. (1988). *Draw A Person: A quantitative scoring system.* San Antonio, TX: Psychological Corporation.
Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A.,Margolis, A., & Velasquez,R. (2004). Psychological testing on the Internet. *American Psychologist, 59,* 150–162.
Naglieri, J. A., Mcneish, T. J., & Bardos, A. N. (1991). *Draw a Person: Screening procedure for emotional disturbance.* San Antonio, TX: Psychological Corporation.
National Association of School Psychologists. (2000). *Principles for professional ethics.* Bethesda, MD: Author.
National Center for Education Statistics. (2001). *The condition of education 2001.* Washington, DC: U.S. Department of Education. (For web copy on pdf. file, go to nces.ed.gov).
Neisser, U. (1998). *The rising curve: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.
Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A.W., Brody,N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77– 101.
Nelson, L. D., Satz, P., & D'elia, L. F. (1994). *Neuropsychology Behavior and Affect Profile.* Palo Alto, CA: Mind Garden.
Nelson, L. D., Satz, P., & D'elia, L. F. (2009). *Neuropsychology Behavior and Affect Profile-D (NBAP-D) Manual (Self and Other Forms).* Menlo Park, CA: Mind Garden.
Nester, M. A. (1994). Psychometric testing and reasonable accommodation for persons with disabilities. In S. M. Bruyere & J. O'Keefe (Eds.), *Implications of the Americans with Disabilities Act for psychology (pp. 25–36)*. New York: Springer Publishing.
Nettelbeck, T. (2011). Basic processes of intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 371–393). New York: Cambridge University Press.
Newmark, C. S., & Mccord, D. M. (1996). The Minnesota Multiphasic Personality Inventory-2 (MMPI-2). In C. S. Newmark (Ed.), *Major psychological assessment instruments* (2nd ed., pp. 1–58). Needham Heights, MA: Allyn & Bacon.
Nichols, D. S. (1992). Review of the Minnesota Multiphasic Personality Inventory-2. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 562–565). Lincoln: University of Nebraska Press.
Nietzel, M. T., Bernstein, D. A., & Milich, R. (1998). *Introduction to clinical psychology* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.
Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130–159.
Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Upper Saddle River, NJ: Pearson.
Norcross, J. C.,Hogan, T. P.,&Koocher, G. P. (2008*). Clinician's guide to evidence based practices: Mental health and the addictions.* New York: Oxford.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
O'connor, M. G., & Kaplan, E. F. (2003). Age-related changes in memory. In J. Demick & C. Andreoletti (Eds.), *Handbook of adult development* (pp. 121–130). New York: Kluwer Academic/Plenum.
Odell, C. W. (1928). *Traditional examinations and new type tests.* New York: Century.
Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbana, IL: University of Illinois Press.
Oswald, D. P. (2005). Review of Piers-Harris Children's Self-Concept Scale. In R. A. Spies, B. S. Plake, & L.

L. Murphy (Eds.), *The sixteenth mental measurements yearbook* (pp. 790–792). Lincoln, NE: Buros Institute of Mental Measurements.
Othmer, E. (2002). *The clinical interview using DSM-IVTR, Volumes 1–2.* Washington, DC: American Psychiatric Press.
Otis, A. S., & Lennon, R. T. (2003). *Otis-Lennon School Ability Test, Eighth Edition, technical manual.* San Antonio, TX: Harcourt Educational Measurement.
Otto, R. K., & Weiner, I. B. (Eds.). (2013). *Handbook of psychology, Volume 11: Forensic psychology* (2nd ed.). Hoboken, NJ: Wiley.
Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Reviews of Psychology, 57,* 401–421.
Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76,* 561–565.
Parents in Action on Special Educ. v. Hannon, 506 F. Supp. 831 (E. D. 111.1980).
Parker, K. C. H., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103,* 367–373.
Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer.
Penry v. Lynaugh, 492 U.S. 302 (1989).
Peterson, C. (2006). *A primer in positive psychology.* New York: Oxford.
Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues.* New York: Oxford University Press.
Phillips, S. E. (Ed.). (2000). Defending a high school graduation test: GI Forum v. Texas Education Agency [Special issue]. *Applied Measurement in Education, 13*(4).
Piaget, J. (1950). *The psychology of intelligence* (M. Piercy & D. E. Berlyne, Trans.). London: Routledge & Paul.
Piaget, J. (1983). Piaget's theory. In P. H. Mussen (Ed.), *Handbook of child psychology,* Vol. I (4th ed., pp. 103–128). New York: Wiley.
Piaget, J., & Inhelder, B. (1969). *The psychology of the child* (H. Weaver, Trans.). New York: Basic Books.
Piers, E. V. (1996). *Piers-Harris Children's Self-Concept Scale, revised manual.* Los Angeles, CA: Western Psychological Services.
Piers, E. V., & Herzberg, D. S. (2002). Piers-Harris Children's Self-Concept Scale: *Manual* (2nd ed.). Los Angeles: Western Psychological Services.
Piotrowski, C. (1996). Use of the Rorschach in forensic practice. *Perceptual and Motor Skills, 82,* 254.
Piotrowski, C. (1999). Assessment practices in the era of managed care: Current status and future directions. *Journal of Clinical Psychology, 55,* 787–796.
Piotrowski, C., & Keller, J. W. (1984). Attitudes toward clinical assessment by members of the AABT. *Psychological Reports, 55,* 831–838.
Piotrowski, C., & Keller, J. W. (1989). Psychological testing in outpatient mental health facilities: A national study. *Professional Psychology: Research and Practice, 20,* 423–425.
Piotrowski, C.,& Lubin, B. (1990). Assessment practices of health psychologists: Survey of APA division 38 clinicians. *Professional Psychology: Research and Practice, 21,* 99–106.
Piotrowski, Z.A. (1937). The Rorschach ink-blot method in organic disturbances of the central nervous system. *Journal of Nervous and Mental Disorders, 86,* 525–537.
Piotrowski, Z. A. (1957). Perceptanalysis. New York: Macmillan.
Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71,* 53–104.
Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40,* 397–404.
Plake, B. S., & Impara, J. C. (1999). *Supplement to the thirteenth mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.
Plomin, R., & Defries, J. C. (1998). The genetics of cognitive abilities and disabilities. *Scientific American, 278*(5), 62–69.

Plomin, R., & Spinath, F. M. (2004). Intelligence: Genetic, genes, and genomics. *Journal of Personality and Social Psychology, 86*(1), 112–129.
Plomin, R., Defries, J. C., Craig, I.W., &Mcguffin, P. (Eds.). (2003). *Behavioral genetics in the postgenomic era.* Washington, DC: American Psychological Association.
Plomin, R.,Defries, J. C.,Mcclearn, G. E.,&Mcguffin, P. (2008). *Behavioral genetics* (5th ed.). New York: Worth.
Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. Educational *Measurement: Issues and Practice, 16*(2), 9–13.
Power,M. J. (2003). Quality of life. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 427–441). Washington, DC: American Psychological Association.
Prifitera, A., & Saklofske, D. H. (Eds.). (1998). *WISCIII clinical use and interpretation: Scientist-practitioner perspectives.* San Diego, CA: Academic Press.
Prifitera, A., Saklofske, D. H., & Weiss, L. G. (Eds.). (2005). *WISC-IV clinical use and interpretation: Scientist-practitioner perspectives.* Boston: Elsevier Academic.
Prifitera, A., Saklofske, D. H., & Weiss, L. G. (Eds.). (2008). *WISC-IV clinical assessment and intervention* (2nd ed.). London, UK: Elsevier.
Psychological Corporation. (1992). *Wechsler Individual Achievement Test Manual.* San Antonio, TX: Author.
Psychological Corporation. (1999). *Wechsler Abbreviated Scale of Intelligence.* San Antonio, TX: Author.
Ramsay, M. C., Reynolds, C. R., & Kamphaus, R. W. (2002). *Essentials of behavioral assessment.* New York: Wiley.
Rapaport, C., Gill, M., & Schafer, J. (1946). *Diagnostic psychological testing* (Vol. 2). Chicago: Year Book Publishers.
Raven, J. C. (1976). *Standard Progressive Matrices.* Oxford, UK: Oxford Psychologists Press.
Raven, J. C., Court, J. H., & Raven, J. (1992). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 3.* Oxford, UK: Oxford Psychologists Press.
Raven, J., Court, J. H., & Raven, J. C. (1993). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 1 general overview.* Oxford, UK: Oxford Psychologists Press.
Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14,* 369–415.
Raymond, M. R. (2002). A practical guide to practice analysis for credentialing examinations. *Educational Measurement: Issues and Practice, 21*(3), 25–37.
Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13–l6.
Reisman, J. M. (1976). *A history of clinical psychology* (enlarged ed.) New York: Irvington.
Reitan, R., & Wolfson, D. (1989). The Seashore Rhythm Test and brain functions. *The Clinical Neuropsychologist, 3,* 70–78.
Reitan, R., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and interpretation.* Tucson, AZ: Neuropsychology Press.
Renaissance Learning. (2003). *STAR Math CS technical manual.* Wisconsin Rapids, WI: Author.
Reschly, D. J. (2000). The present and future status of school psychology in the United States. *School Psychology Review, 29,* 507–522.
Retzlaff, P. (1998). Review of the Millon Clinical Multiaxial Inventory—III. In B. S. Plake & J. C. Impara (Eds.), *The thirteenth mental measurements yearbook* (pp. 667–668). Lincoln: University of Nebraska Press.
Reynolds, C. R. (1994). Bias in testing. In R. J. Stern-Berg (Ed.), *Encyclopedia of human intelligence* (pp. 175–178). New York: Macmillan.
Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children* (2nd ed.). Circle Pines, MN: AGS.
Reynolds, C. R., & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology. Vol 10: Assessment psychology* (pp. 67–93). New York: Wiley.
Ricci V. Destefano, 129 S. Ct. 2658, 2671, 174 L. Ed. 2d 490 (2009).
Riverside Publishing. (2001). *2001 assessment catalog.* Itasca, IL: Author.

Robertson, G. J. (1986). Establishing test purchaser qualifications. In R. B. Most (Ed.), Test purchaser qualifications: Present practice, professional needs, and a proposed system. *Issues in Scientific Psychology.* Washington, DC: American Psychological Association, Scientific Affairs Office.

Robertson, G. J. (n.d.). Test Service Notebook 30—Innovation in the assessment of individual differences: Development of the first group mental ability test. New York: Harcourt Brace Jovanovich.

Robinson, J. P., Shaver, P. R., & Wrightsman, L.S. (Eds.) (1991). Measures *of personality and social psychological attitudes.* San Diego: Academic Press.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity technical adequacy, and implications* (pp. 213–231). Mahwah, NJ: Erlbaum.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40,* 163–184.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24* (2), 3–13.

Rogers, R. (2001). Handbook of diagnostic and structured interviewing. New York: Guilford.

Roid, G. H. (2003a). *Stanford-Binet Intelligence Scales, Fifth Edition, examiner's manual.* Itasca, IL: Riverside.

Roid, G. H. (2003b). *Stanford-Binet Intelligence Scales, Fifth Edition, technical manual.* Itasca, IL: Riverside.

Rorschach, H. (1921). *Psychodiagnostik*. Berne, Switzerland: Bircher.

Rotter, J. B., & Rafferty, J. E. (1950). *The Rotter Incomplete Sentences Blank*. New York: Psychological Corporation.

Rotter, J. B., Lah, M. I., & Rafferty, J. E. (1992). *Manual: Rotter Incomplete Sentences Blank* (2nd ed.). San Antonio, TX: Psychological Corporation.

Routh, D. K., & Reisman, J. M. (2003). Clinical psychology. In D. Freedheim (Ed.), *Handbook of psychology, history of psychology, Vol. 1* (pp. 337–355). Hoboken, NJ: Wiley.

Ruch, G. M. (1924). *The improvement of the written examination.* Chicago, IL: Scott, Foresman and Company.

Ruch, G. M. (1929). *The objective or new-type examination: An introduction to educational measurement.* Chicago, IL: Scott, Foresman and Company.

Ruch, G. M., & Rice, G. A. (1930). *Specimen objective examination.* [A collection of examinations awarded prizes in a national contest in the construction of objective or new-type examinations, 1927–1928.] Chicago, IL: Scott, Foresman and Company.

Ruch, G. M., & Stoddard, G. D. (1927). *Tests and measurements in high school instruction.* Yonkers, NY: World Book.

Runners World. (1999). The human race. *Runners World, 34*(9), 123.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11,* 235–294.

Russell, E.W. (1992). Reliability of the Halstead Impairment Index: A simulation and reanalysis of Matarazzo et al. (1974). *Neuropsychology, 6,* 251–259.

Russell, E. W. (1995). The accuracy of automated and clinical detection of brain damage and lateralization in neuropsychology. *Neuropsychology Review, 5,* 1–68.

Ryan, R.M. (1985). Thematic Apperception Test. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques, vol. II,* (pp. 799–814). Kansas City, MO: Test Corporation of America.

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology, 52,* 141–166.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112–118.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin, 135,* 1–22.

Sandoval, J., & Miille, P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology, 48,* 249–253.

Sandoval, J. (2003) Review of the Woodcock-Johnson(r) III. In B. S. Plake, J. C. Impara, & R. A. Spies

(Eds.), *The fifteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
Sands, W. A., & Waters, B. K. (1997). Introduction to ASVAB and CAT. InW. A. Sands, B. K.Waters, & J.R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 3–10).Washington, DC: American Psychological Association.
Sands, W.A., Waters, B.K., & Mcbride, J.R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.
Saucier, G. (2009). Recurrent personality dimensions in inclusive lexical studies: Indications for a big six structure. *Journal of Personality, 77,* 1577–1614.
Schaeffer, G. A., Briel, J. B., & Fowles, M. E. (2001). *Psychometric evaluation of the new GRE Writing Assessment: ETS Research Report 01-08.* Princeton, NJ: Author.
Scheerenberger, R. C. (1987). *A history of mental retardation: A quarter century of promise.* Baltimore: P. H. Brookes.
Schinke, S. (1995). Review of the Eating Disorder Inventory- 2. In J. C. Conoley & J. C. Impara (Eds.), *Twelfth mental measurements yearbook* (p. 335). Lincoln: University of Nebraska Press.
Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162–173.
Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods, 8*(2), 206–224.
Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43,* 627–670.
Schmitt, K. (1995). What is licensure? In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 3–32). Lincoln, NE: Buros Institute of Mental Measurements.
Searls, E. F. (1997). How to detect reading/learning disabilities using the WISC-III. Newark, DE: International Reading Association.
Seashore, H. G. (n.d.). *Test service notebook 148: Methods of expressing test scores.* San Antonio, TX: Psychological Corporation.
Seddon, G. M. (1978). The properties of Bloom's Taxonomy of Educational Objectives for the cognitive domain. *Review of Educational Research, 48*(2), 303–323.
Seligman, M. E. P., Steen, T. A., Park, N., & Peterson, C. (2005). Positive psychology progress: Empirical validation of interventions. *American Psychologist, 60,* 410–421.
Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist, 55,* 5–14.
Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology, 62*(1), 123–146.
Shavelson, R. J., & Webb, N. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.
Shavelson, R. J.,Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44,* 922–932.
Shaw, M. E., & Wright, J. M. (1967). Scales *for the measurement of attitudes.* New York: McGraw-Hill.
Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8.
Sherer, M., Parsons, O. A., Nixon, S., & Adams, R. L. (1991). Clinical validity of the Speech-Sounds Perception Test and the Seashore Rhythm Test. *Journal of Clinical and Experimental Neuropsychology, 13,* 741–751.
Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Erlbaum.
Shermis, M. D., & Daniels, K. E. (2003). Norming and scoring for automated essay scoring. In J. Burstein & M. D. Shermis (Eds.), Automated essay scoring: A cross-disciplinary *perspective* (pp. 169–180). Mahwah, NJ: Erlbaum.
Shrout, P. E.,&Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.
Shuster, E. (1997). Fifty years later: The significance of the Nuremberg Code. *The New England Journal of*

Medicine, 337, 1436–1440.
Sireci, S.G., Li, S.,&Scarpati, S. (2006). *The Effects of Test Accommodation on Test Performance: A Review of the Literature*. Center for Educational Assessment, Research Report No. 485. Amherst, MA: School of Education, University of Massachusetts Amherst.
Slora, K. B. (1991). An empirical approach to determining employee deviance base rates. In J. W. Jones (Ed.), *Preemployment honesty testing: Current research and future directions* (pp. 21–38). New York: Quorum.
Smith, C. P. (Ed.). (1992). *Motivation and personality: Handbook of thematic content analysis.* New York: Cambridge University Press.
Smith, J. D. (1997). Mental retardation: Defining a social invention. In R. L. Taylor (Ed.), *Assessment of individuals with mental retardation* (pp. 3–12). San Diego, CA: Singular Publishing Group.
Smith, S. R., Gorske, T. T., Wiggins, C., & Little, J. A. (2010). Personality assessment use by clinical neuropsychologists. *International Journal of Testing, 10,* 6–20.
Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist, 42,* 137–144.
Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Washington, DC: Author.
Spangler, W.D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin, 112,* 140–154.
Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales, Interview Edition Expanded Form manual.* Circle Pines, MN: American Guidance Service.
Sparrow, S. S., Cichetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales* (2nd ed.), Survey Forms manual. Circle Pines, MN: American Guidance Service.
Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–292.
Spearman, C. (1927a). *The abilities of man: Their nature and measurement.* New York: Macmillan.
Spearman, C. (1927b). *The nature of "intelligence" and the principles of cognition* (2nd ed.). London: Macmillan.
Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? *American Psychologist, 60,* 950–958.
Spielberger, C. D. (1973). *Manual for the State-Trait Anxiety Inventory for Children.* Palo Alto, CA: Consulting Psychologists Press.
Spielberger, C. D. (1983). *State-Trait Anxiety Inventory for Adults sampler set: Manual, test, scoring key.* Redwood City, CA: Mind Garden.
Spies, R. A., & Plake, B. S. (2005). *The sixteenth mental measurements yearbook*. Lincoln. University of Nebraska Press.
Spies, R. A., Carlson, J. F., & Geisinger, K. F. (2010). *The eighteenth mental measurements yearbook*. Lincoln, NE: Buros Center for Testing.
Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary.* New York: Oxford University Press.
Sternberg, R. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge, UK: Cambridge University Press.
Sternberg, R. J. (Ed.). (1994a). *Encyclopedia of human intelligence.* New York: Macmillan.
Sternberg, R. J. (1994b). Triarchic theory of human intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 1087–1091). New York: Macmillan.
Sternberg, R. J. (2011). The theory of successful intelligence. In J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 504–527). New York: Cambridge University Press.
Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence?* Norwood, NJ: Ablex.
Sternberg, R. J., & Grigorenko, E. (Eds.). (1997). *Intelligence, heredity, and environment.* Cambridge, UK: Cambridge University Press.
Sternberg, R. J., & Kaufman, J. C. (1998). Human abilities. *Annual Review of Psychology, 49,* 479–502.
Sternberg, R. J., & Kaufman, S. B. (2011). The *Cambridge handbook of intelligence.* New York: Cambridge University Press.

Sternberg, R. J., &Wagner, R. K. (Eds.). (1986). *Practical intelligence: Nature and origins of competence in the everyday world.* Cambridge: Cambridge University Press.
Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley.
Stewart, T. M., & Williamson, D. A. (2004). Assessment of eating disorders. In M. Hersen (Ed.), *Psychological assessment in clinical practice: A pragmatic guide* (pp. 175–195). New York: Bruner-Routledge.
Stokes, T. L.,&Bohrs, D. A. (2001). The development of a same-different inspection time paradigm and the effects of practice. *Intelligence, 29,* 247–261.
Strack, S. (2008). *Essentials of Mill on inventories assessment* (3rd ed.). New York: Wiley.
Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York: Oxford.
Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.
Suzuki, L. A., & GUTKIN, T. B. (1994a). Asian Americans. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 140–144). New York: Macmillan.
Suzuki, L. A., & Gutkin, T. B. (1994b). Hispanics. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 539–545). New York: Macmillan.
Suzuki, L. A., & Gutkin, T. B. (1994c). Japanese. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 625–629). New York: Macmillan.
Suzuki, L. A., Short, E. L., & Lee, C. S. (2011). Racial and ethnic differences in intelligence in the United States. In J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 273–292). New York: Cambridge University Press.
Swartz, J. D. (1978). Review of the Thematic Apperception Test. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 1127–1130). Lincoln: University of Nebraska Press.
Sweet, J. J., Meyer, D. G., Nelson, N.W., & Moberg, P. J. (2011). The TCN/AACN 2010 "Salary Survey": Professional practices, beliefs, and incomes of U.S. neuropsychologists. *The Clinical Neuropsychologist, 25,* 12–61.
Sweet, J. J.,Moberg, P. J., & Suchy, Y. (2000). Ten-year follow-up survey of clinical neuropsychologists: Part I. Practices and beliefs. *The Clinical Neuropsychologist, 14,* 18–37.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.
Tabachnik, B.G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York: Harper Collins.
Taleporos, E. (1998). Consequential validity. A practitioner's perspective. *Educational Measurement: Issues and Practice, 17*(2), 20–23.
Tarasoff. Univ. Of Ca., P. 2d 334 (9th Cir. 1976). TAYLOR, H. C., & RUSSELL, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.
Tellegen, A., & Ben-Porath, Y. S. (2008). *MMPI-2-RF* (Minnesota Multiphasic Inventory-2-Restructured Form) technical manual. Minneapolis, MN: University of Minnesota.
Tellegen, A., Ben-Porath, Y. S., Mcnulty, J. L.,Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The MMPI-2 Restructured Clinical Scales: Development, validation, and interpretation.* Minneapolis, MN: University of Minnesota.
Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length big five and big six personality questionnaires. *Psychological Assessment, 23,* 995–1009.
Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
Thompson, A. P., Lobello, S. G., Atkinson, L., & Chisolm, V. (2004). Brief intelligence testing in Australia, Canada, the United Kingdom, and the United States. *Professional Psychology: Research and Practice, 35,* 286–290.
Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota,

National Center on Educational Outcomes. Retrieved September 26, 2012, from http://education.umn.edu/NCEO/OnlinePubs/ Technical34.htm

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: Teachers College, Columbia University.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Thorndike, R. L., Hagen, E. P., & Sattler, J.M. (1986). *Stanford-Binet Intelligence Scale, Fourth Edition.* Itasca, IL: Riverside Publishing.

Thurlow, M.L., & Ysseldyke, J. E. (2001). Standard setting challenges for special populations. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 387–409). Mahwah, NJ: Lawrence Erlbaum.

Thurlow, M. L., & Ysseldyke, J. E. (2002). *Including students with disabilities in assessments.* Washington, DC: National Education Association.

Thurlow,M. L., Elliott, J. L., & Ysseldyke, J. E. (1998). *Testing students with disabilities: Practical strategies for complying with district and state requirements.* Thousand Oaks, CA: Corwin.

Thurstone, L. L. (1935). *The vectors of the mind.* Chicago: University of Chicago Press.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago: University of Chicago Press.

Thurstone, L. L. (1959*). The measurement of values.* Chicago: University of Chicago Press.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church.* Chicago: University of Chicago Press.

Tinsley, H. E. (1994). The NEO Personality Inventory—Revised. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques, vol. X* (pp. 443–456). Kansas City, MO: Test Corporation of America.

Todd, J., & Bohart, A. C. (1994). *Foundations of clinical and counseling psychology* (2nd ed.). New York: HarperCollins.

Tombaugh, T. N. (1996). Test of Memory Malingering (TOMM). North Tonawanda, NY: Multi-Health Systems.

Tombaugh, T.N., &Mcintyre, N. J. (1992). The Mini-Mental State Examination: A comprehensive review. *Journal of the American Geriatrics Society, 40,* 922–935.

Toops, H. A. (1921). *Trade tests in education.* [Teachers College Contributions to Education No. 115.] New York: Teachers College, Columbia University.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & C. W. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29–44). Hillsdale, NJ: Erlbaum.

Trull, T. J., & Prinstein, M. J. (2013). *Clinical psychology* (8th ed.). Belmont, CA: Wadsworth.

Tryon, W. (2008). History and theoretical foundations. In M. Hersen & A. Gross (Eds.), *Handbook of clinical psychology, Vol. 1: Adults* (pp. 3–37). Hoboken, NJ: Wiley.

Turkheimer, E. (1994). Socioeconomic status and intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 992–1000). New York: Macmillan.

Turner, S. M., Demers, S. T., Fox, H. R., & Reed, G. M. (2001). APA's guidelines for test user qualifications: An executive summary. American Psychologist, 56, 1099–1113.

U.S. Government Printing Office. (1949). *Trials of war criminals before the Nuremberg military tribunals under control council law No. 10, Vol. 2* (pp. 181–182). Washington, DC: Author.

Umberger, F. G. (1985). Peabody Picture Vocabulary Test—Revised. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques, Vol. III* (pp. 488–495). Kansas City, MO: Test Corporation of America.

United States Department of Health, Education, and Welfare. (1979). *The Belmont report.* Washington, DC: U.S. Government Printing Office.

United States Department of Labor. (1991). *Dictionary of occupational titles* (4th ed., rev.). Washington, DC: U.S. Government Printing Office.

van der linden, w. j., & glas, c. a. w. (Eds.). (2010). *Elements of adaptive testing.* New York: Springer.

Vandenberg, G. H. (1993). *Court testimony in mental health: A guide for mental health professionals and attorneys.* Springfield, IL: Charles C Thomas.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature:

Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70.
Vernon, P. A. (1984). Advanced Progressive Matrices. In D. Keyser & R. Sweetland (Eds.), *Test critiques,* Vol. I (pp. 47–50). Austin, TX: PRO-ED.
Vernon, P. E. (1947). Research on personnel selection in the Royal Navy and the British Army. *American Psychologist, 2,* 35–51.
Vernon, P. E. (1950). *The structure of human abilities.* London: Methuen.
Vernon, P. E. (1961). *The structure of human abilities* (2nd ed.). London: Methuen.
Vernon, P. E. (1965). Ability factors and environmental influences. *American Psychologist, 20,* 723–733.
Viglione, D. J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment, 11*(3), 251–265.
Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and futures. Psychological Assessment, 13, 452–471.
Volpe, R. J., & Dupaul, G. J. (2001). Assessment with brief behavior rating scales. In J. J. W. Andrews, D. H. Saklofske, & H. L. Janzen (Eds.), *Handbook of psychoeducational assessment: Ability, achievement, and behavior in children* (pp. 357–385). San Diego, CA: Academic Press.
Vraniak, D. A. (1994). Native Americans. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 747–754). New York: Macmillan.
Wade, T. C., & BAKER, T. B. (1977). Opinions and use of psychological tests: A survey of clinical psychologists. *American Psychologist, 32,* 874–882.
Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
Walsh, W. B., & Osipow, S.H. (Eds.). *Handbook of vocational psychology* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
Wasyliw, O. E. (2001). Review of the Peabody Picture Vocabulary Test-III. In B. S. Plake & J. C. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 909–911). Lincoln: University of Nebraska Press.
Watkins, C. E., Campbell, V. L., & Mcgregor, P. (1988). Counseling psychologists' uses of and opinions about psychological tests: A contemporary perspective. *The Counseling Psychologist, 16,* 476–486.
Watkins, C. E., Campbell, V. L., & Nieberding, R. (1994). The practice of vocational assessment by counseling psychologists. *The Counseling Psychologist, 22*(1), 115–128.
Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26,* 54–60.
Watterson, B. (1988). Something under the bed is drooling. Kansas City, MO: Andrews and McMeel.
Webster, R. E. (1994). Woodcock-Johnson Psycho-Educational Battery—Revised. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques, vol. X* (pp. 804–815). Austin, TX: PRO-ED.
Wechsler, D. (1949). *Wechsler Intelligence Scale for Children: Manual.* New York: Psychological Corporation.
Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Williams and Wilkins.
Wechsler, D. (1974). *The collected papers of David Wechsler.* New York: Academic Press.
Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third Edition, manual.* San Antonio, TX: The Psychological Corporation.
Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale—Third Edition: Administration and scoring manual.* San Antonio, TX: The Psychological Corporation.
Wechsler, D. (1997b). *Wechsler Adult Intelligence Scale—Third Edition, Wechsler Memory Scale—Third Edition: Technical manual.* San Antonio, TX: The Psychological Corporation.
Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children— Fourth Edition: Administration and scoring manual.* San Antonio, TX: Psychological Corporation.
Wechsler, D. (2003b). Wechsler Intelligence Scale for Children—Fourth Edition: Technical and interpretive manual. San Antonio, TX: Psychological Corporation.
Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale—Fourth Edition: Technical and interpretive manual.* San Antonio, TX: Pearson.
Wechsler, D. (2008b). *Wechsler Adult Intelligence Scale—Fourth Edition: Administration and scoring manual.*

San Antonio, TX: Pearson.
Wechsler, D. (2009a). *Wechsler Memory Scale—Fourth Edition: Administration and scoring manual.* San Antonio, TX: NCS Pearson.
Wechsler, D. (2009b). *Wechsler Memory Scale—Fourth Edition: Technical and interpretive manual.* San Antonio, TX: NCS Pearson.
Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence* (2nd ed.). San Antonio, TX: Pearson.
Weiner, I. B. (2001). Advancing the science of psychological assessment: The Rorschach inkblot method as exemplar. *Psychological Assessment, 13,* 423–432.
Weiner, I. B. (2003). *Principles of Rorschach interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.
Weiner, I. B., & Greene, R. L. (2008). *Handbook of personality assessment.* Hoboken, NJ: Wiley.
Weiss, L. G., Saklofske, D. H., Coalson, D., & Raiford, S. E. (Eds.). (2010). *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives.* London, UK: Elsevier.
WELSH, J. R., Kucinkas, S. K., & Curran, L. T. (1990). Armed Services Vocational Battery (ASVAB): *Integrative Review of Validity Studies.* Brooks Air Force Base, TX: AIR Force Systems Command.
West Publishing. (1984). *The guide to American law.* St. Paul, MN: Author.
West Publishing. (1990). *United States code annotated, Title 20, Education, §1241 to 3400.* St. Paul, MN: Author.
Whitfield, E. A., Feller, R.W., & Wood, C. (2008). *A counselor's guide to career assessment instruments* (5th ed.). Broken Arrow, OK: National Career Development Association.
Widaman, K. F., & Mcgrew, K. S. (1996). The structure of adaptive behavior. In J. W. Jacobson & J. A. Mulick (Eds.), *Manual of diagnosis and professional practice in mental retardation* (pp. 197–210). Washington, DC: American Psychological Association.
Widiger, T. A. (2001). Review of the Millon Clinical Multiaxial Inventory—III. In J. C. Impara & B. S. Plake (Eds.), *The fourteenth mental measurements yearbook* (pp. 767–769). Lincoln: University of Nebraska Press.
Widiger, T. A., & Costa, P. T. (Eds.). (2012). *Personality disorders and the five-factor model of personality* (3rd ed.). Washington, DC: American Psychological Association.
Wiger, D. E. (2002). *Essentials of interviewing.* New York: Wiley.
Wiggins, J. S., & Trapnell, P. D. (1997). Personality structure: The return of the Big Five. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 737–766). San Diego, CA: Academic Press.
Williams, J. M., Mathews, A., & Macleod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*(1), 3–24.
Williams, K. T.,& Wang, J. J. (1997). *Technical references to the Peabody Picture Vocabulary Test—Third Edition (PPVT—III).* Circle Pines, MN: American Guidance Service.
Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). "Mental model" comparison of automated and human scoring. *Journal of Educational Measurement, 36,* 158– 184.
Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people.* Needham Heights, MA: Allyn & Bacon.
Willse, J. T. (2010). Review of the Wechsler Individual Achievement Test-Third Edition. In Spies, R. A., Carlson, J. F., & Geisinger, K. F. (Eds.), *The eighteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
Winer, B. J. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.
Wolfe, J. H., Moreno, K. E., & Segall, D. O. (1997). Evaluating the predictive validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 175–180). Washington, DC: American Psychological Association.
Wolman,B.B. (Ed.). (1985). *Handbook of intelligence: Theories, measurements, and applications.* New York: Wiley.
Wonderlic. (2002). *Wonderlic Personnel Test and Scholastic Level Exam user's manual.* Libertyville, IL: Author.
Wonderlic. (2007). *Wonderlic Personnel Test-Revised: Administrator's manual.* Libertyville, IL: Author.
Wood, B. D. (1923). *Measurement in higher education.* Yonkers, NY: World Book.
Wood, B. D. (1927). *New York experiments with new-type modern language tests.* [Publications of the American

and Canadian Committees on Modern Languages.] New York: Macmillan.
Woodcock, R. W., Mcgrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III.* Itasca, IL: Riverside Publishing.
World Health Organization. (1996). *ICD-10 guide for mental retardation.* Geneva, Switzerland: Author.
Wresch, W. (1993). The imminence of grading essays by computer—25 years later. *Computers and Composition, 10,* 45–58.
Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16,* 33–45, 52.
Wrightsman, L. S. (1991). *Psychology and the legal system* (2nd ed.). Pacific Grove, CA: Brooks/Cole.
Zieky,M. J. (2001). So much has changed: How the setting of cutscores has changed since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51). Mahwah, NJ: Lawrence Erlbaum.
Zytowski, D. G. (1992). Three generations: The continuing evolution of Frederic Kuder's interest inventories. *Journal of Counseling and Development, 71*(2), 245–248.
Zytowski, D. G. (2005). *Kuder Career Search with Person Match: Technical manual version 1.1.* Adel, IA: National Career Assessment Services.

Índice