Introducción a la Psicometría

Teoría clásica y TRI

José Muñiz

PIRÁMIDE

Introducción a la Psicometría

Teoría clásica y TRI

JOSÉ MUÑIZ
CATEDRÁTICO DE PSICOMETRÍA DE LA UNIVERSIDAD DE OVIEDO

Introducción a la Psicometría

Teoría clásica y TRI

EDICIONES PIRÁMIDE

COLECCIÓN «PSICOLOGÍA»

Director:
Francisco J. Labrador
Catedrático de Modificación de Conducta
de la Universidad Complutense de Madrid

Edición en versión digital

Para cualquier información pueden dirigirse a piramide_legal@anaya.es
Juan Ignacio Luca de Tena, 15. 28027 Madrid
Teléfono: 91 393 89 89
www.edicionespiramide.es
ISBN digital: 978-84-368-3933-3

A mis hijas Laura y Marta.
A mi mujer Alejandra.

Índice

Prólogo

Los test son instrumentos de medida que utilizan habitualmente los psicólogos para obtener información que les ayude a tomar decisiones bien fundamentadas. Estas decisiones pueden tener repercusiones importantes en la vida de las personas; por tanto, es clave que los test utilizados cumplan unos requisitos de calidad psicométrica demostrables. Precisamente de eso trata el libro que tiene entre las manos, de mostrar cuáles deben ser las propiedades psicométricas de los test para que puedan ser utilizados con garantía por los psicólogos. Queridos lectores, lo que van a encontrar en las páginas que siguen es una introducción al campo de la psicometría, como bien anuncia, sin sorpresas, el título del libro. Está pensando para alguien que aún no sabe, pero que pretende saber; por eso la filosofía del libro es iniciar a los estudiantes, profesionales y en general personas interesadas en la evaluación psicométrica. Los contenidos se ajustan a la materia que los estudiantes de grado de psicología, educación y ciencias sociales y de la salud deben aprender en un semestre dedicado a la psicometría. También los ya iniciados pueden repasar y actualizar sus conocimientos, pues nunca viene mal volver a hollar los caminos andados y descubrir detalles que se nos habían pasado desapercibidos. A todos vosotros va dirigido este libro introductorio, que pretende ser de amigable lectura, pues no supone grandes conocimientos estadísticos o psicométricos previos. Si lo leyereis con provecho, os abrirá las puertas a otros textos más avanzados, de los cuales hay actualmente abundancia tanto en nuestra propia lengua española como en inglés, y que se irán citando en su momento para animar a profundizar en los temas tratados. El grueso del libro procede de la fusión y actualización de otros dos previos del autor, publicados en esta misma editorial Pirámide, uno sobre teoría clásica y otro sobre teoría de respuesta a los ítems. Esta fusión ha permitido reunir lo esencial de ambos en un solo volumen, evitando redundancias y actualizando lo hecho, aparte de incluir nuevos temas y suprimir otros. Del enfoque clásico se expone lo fundamental: el modelo, la fiabilidad, la validez, el análisis de los ítems, la transformación de las puntuaciones y la equiparación; sin olvidarnos de los test referidos al criterio y la teoría de la generalizabilidad. En cuanto a la teoría de respuesta a los ítems, tras exponer su lógica y los avances que supone sobre el enfoque clásico, se presentan los principales modelos, la curva característica del test, las funciones de información de los ítems y del test, los bancos de ítems, el funcionamiento diferencial y los test adaptativos informatizados. Finaliza el libro con tres apartados de gran interés aplicado, como son las fases o etapas para construir un test, los problemas implicados en la utilización práctica de los test y unas reflexiones finales sobre el futuro de la evaluación psicométrica. La idea que inspira todo el libro es dar una visión comprensiva, no especializada pero rigurosa, sobre el estado actual de la evaluación psicométrica, sin la cual no es posible un ejercicio profesional basado en evidencias, fieles a la idea de que una evaluación rigurosa es la base de un diagnóstico preciso que a su vez permita una intervención eficaz. Si falla la evaluación, todo lo demás se viene abajo.

El libro no hubiese sido posible sin la ayuda de tantas y tantas personas, unas de forma explícita y otras implícita, que tanto monta. En primer lugar, la fuerza motriz del libro son los alumnos de muchas

partes del mundo a los que vengo impartiendo psicometría desde hace ahora cuarenta años, con especial mención para los de las tres universidades españolas en las que tuve el honor de trabajar: Complutense de Madrid, Baleares y Oviedo. Son ellos quienes con sus dudas y preguntas incisivas me ayudan y estimulan a tratar de presentar la materia de forma amena y comprensible. No sé si lo habré logrado. Muchas gracias también a los compañeros de nuestro grupo de investigación de psicometría: Marcelino Cuesta, Yolanda de la Roca, Rubén Fernández, Eduardo Fonseca, Eduardo García-Cueto, Elena Govorova, Luis Manuel Lozano, Teresa Martínez, Víctor Martínez-Loredo, Fernando Menéndez, Ignacio Pedrosa, Elsa Peña, Francisco Prieto, Javier Suárez y Pamela Woitschach. Nuestras sesiones y discusiones para sacar adelante los proyectos constituyen para mí una verdadera fuente de aprendizaje y motivación. He tenido muchos maestros, caminamos a hombros de gigantes, pero no quiero dejar de citar a dos de quienes más he aprendido directamente en muchos sentidos: Mariano Yela y Ronald Hambleton. Sus profundos conocimientos psicométricos, unidos a su talante personal y bonhomía, fueron una bendición para mí; muchas gracias, maestros. Nada saldría adelante sin el apoyo decidido de la directora editorial Inmaculada Jorge, que tuvo claro el proyecto desde el principio y me animó a ello. Se agradece sinceramente. Por supuesto, la familia siempre está al quite: a mis hijas Laura y Marta y a mi mujer Alejandra va dedicado el libro. Perdonad todos los demás no citados, sé que estáis ahí, como bien nos enseñó Bertolt Brecht en su bello poema «Preguntas de un obrero ante un libro». Muchas gracias a todos por todo. Espero que el libro resulte de interés y ayude a penetrar en los arcanos de la psicometría, pues de eso se trata.

1 Introducción

La psicometría puede definirse en términos generales como el conjunto de métodos, técnicas y teorías implicadas en la medición de las variables psicológicas. Como su nombre indica, trataría de todo aquello relacionado con la medición de lo psicológico. Ahora bien, de la medición de lo psicológico se ocupa también cualquier otro acercamiento riguroso al estudio del comportamiento humano; lo que constituiría lo específico de la psicometría sería su énfasis y especialización en aquellas propiedades métricas exigibles a las mediciones psicológicas independientemente del campo sustantivo de aplicación y de los instrumentos utilizados. Así, por ejemplo, aspectos como la fiabilidad o la validez de las mediciones, por citar dos de los más conocidos, constituyen requisitos exigibles para cualquier evaluación psicológica, sea cual fuere su ámbito de aplicación y enfoque. Este tipo de especialización de la psicometría en las categorías métricas que atraviesan los distintos campos sustantivos de la psicología da lugar a que los tipos de contenidos sobre los que trabajan los psicómetras resulten bastante amplios y variados.

Una ojeada a los congresos organizados por las sociedades psicométricas europea y americana, así como a las revistas científicas del área, permite articular la mayor parte de la temática psicométrica en torno a cinco grandes bloques: teoría de la medición, que abarcaría todo lo relativo a la fundamentación teórica de la medida; teoría de los test, donde se explicitan la lógica y los modelos matemáticos subyacentes a la construcción y uso de los test; escalamiento psicológico, que aborda la problemática inherente al escalamiento de estímulos psicológicos; escalamiento psicofisico, que hace lo propio con los estímulos fisicos, y técnicas multivariadas, que junto con el resto de tecnología estadística resultan imprescindibles para la construcción y análisis de los instrumentos de medida.

Como se puede observar, el campo de referencia del término «psicometría» es amplísimo, y no es nuestro objetivo entrar aquí a realizar una descripción o definición precisa de los bloques citados, cada uno de los cuales está a su vez altamente especializado y estructurado en subáreas. En este libro nos centraremos exclusivamente en la teoría de los test, abordando los dos grandes enfoques, el clásico y la teoría de respuesta a los ítems (TRI). Solo subrayar que el término «psicometría» es mucho más genérico y amplio que el de teoría de los test, con el que erróneamente se le identifica a menudo, tomando la parte por el todo. A partir de los años sesenta se populariza también el término afín de «psicología matemática», utilizado para denominar aquellos trabajos caracterizados por un acercamiento formalizado a los problemas psicológicos, lo cual es básicamente coincidente con lo que se entendía por psicometría. De hecho, Thurstone (1937) utiliza el término «psicología matemática» para caracterizar en pocas palabras el objeto de la sociedad psicométrica americana por él fundada. Lo más específico y diferencial de la psicología matemática respecto de la psicometría serán los modelos matemáticos elaborados para áreas específicas de la psicología, tales como aprendizaje, memoria, percepción, lenguaje, pensamiento, interacción social, etc., que proliferan a partir de los años cincuenta y sesenta, y que son en realidad la razón fundamental que da sentido y apoya la nueva denominación de psicología matemática diferenciada de

la anterior de psicometría. Como ocurriera antes en psicometría, se publican textos con esa denominación genérica (Atkinson, 1964; Coombs, Dawes y Tversky, 1970; Krantz, Atkinson, Luce y Suppes, 1974; Laming, 1973; Luce, Bush y Galanter, 1963; Restle y Greeno, 1970), que luego desaparecerá para dar paso a las subáreas especializadas, aparecen sociedades científicas de psicólogos matemáticos con reuniones y congresos propios desde 1967 y se publican revistas, con mención especial para el *Journal of Mathematical Psychology* y *British Journal of Mathematical and Statistical Psychology.* Que ambos términos, «psicometría» y «psicología matemática», denotan campos muy solapados se comprueba empíricamente al observar que los especialistas publican sus trabajos en las revistas de uno y otro campo y asisten indistintamente a los congresos y reuniones de las respectivas sociedades científicas. Considerar, como hacen Estes (1975) o Greeno (1980), la psicometría como la parte de la psicología matemática dedicada a todo lo relacionado con la medida es tan razonable como considerar la psicología matemática aquella rama de la psicometría dedicada a los modelos matemáticos de procesos psicológicos, pero, eso sí, aquella resulta menos respetuosa con la historia de la psicología, ya que fue «psicometría» el término que originariamente se utilizó para tales menesteres. Así, Thurstone (1937), en su conferencia para la primera reunión anual de la sociedad psicométrica que tuvo lugar en 1936, señala como objetivo de la sociedad «estimular el desarrollo de la psicología como ciencia cuantitativa y racional. O lo que más brevemente puede denominarse psicología matemática». Añade, además, algo que se olvida demasiado a menudo: «A la larga seremos juzgados por la significación, relevancia y consistencia de los principios psicológicos que descubramos», dejando claro desde el principio que la formalización y matematización propias del enfoque psicométrico están al servicio de los problemas psicológicos que tratan de resolver, pero no constituyen un fin en sí mismo para la psicometría.

Quede ahí este breve apunte terminológico para evitar confusiones y desasosiegos al lector que los topare en su deambular por los reales de la psicología, aunque no lo hará con mucha frecuencia, pues, como ya se ha señalado, debido a su amplitud, se han vuelto ambos demasiado genéricos y lo habitual será encontrarse con las denominaciones específicas de las distintas subáreas. Como modesto ejemplo, en este libro se tratará de las teorías de los test, tanto el enfoque clásico como la teoría de respuesta a los ítems.

1. ORÍGENES Y DESARROLLO DE LA TEORÍA DE LOS TEST

El nacimiento formal de la teoría de los test puede ubicarse en los primeros trabajos de Spearman (1904, 1907, 1913), en los que establece los fundamentos de la teoría clásica de los test (TCT). El objetivo central era encontrar un modelo estadístico que fundamentase adecuadamente las puntuaciones de los test y permitiera la estimación de los errores de medida asociados a todo proceso de medición. El modelo lineal clásico propuesto por Spearman, que luego se abordará con cierto detalle, destaca por su sencillez matemática y enjundia psicológica, lo que le garantizará una larga vida. Asume que la puntuación empírica de una persona en un test (X) consta de dos componentes aditivos: uno, la «verdadera» puntuación de la persona en el test (V), y otro, el error (e) que inevitablemente va asociado a todo proceso de medición; es decir, según el modelo:

$$X = V + e$$

A partir de este modelo y unas asunciones mínimas, la teoría clásica desarrollará todo un conjunto de deducciones encaminadas a estimar la cuantía del error que afecta a las puntuaciones de los test. Los inicios fueron como siempre inseguros, pues no era fácil abrirse paso en una psicología poco dada a veleidades cuantitativas, pero los nuevos enfoques se impusieron con rapidez y la mayoría de las universidades incluyeron cursos de teoría de los test en el currículum de sus licenciados. Cuenta Joncich (1968) en la biografía de E. L. Thorndike que cuando este envió una copia de su libro pionero sobre medición (Thorndike, 1904) a su antiguo profesor William James, incluyó una nota diciéndole que obligase a leerlo a todos sus investigadores, pero que no se le ocurriese abrirlo a él, pues las figuras, curvas y fórmulas que contenía le volverían loco. Ello puede dar una idea de la acogida que se esperaba por parte de la

psicología dominante de la época. Los años siguientes conocieron una actividad psicométrica frenética tanto en el campo de la teoría como en la construcción y tecnología de los test, así como en el campo del escalamiento psicofisico y psicológico (Thurstone, 1927a, 1928b; Thurstone y Chave, 1929), muy cercanos por entonces al ámbito de los test. Guilford (1936) tratará de sintetizar en su clásico manual *Psychometric Methods* lo fundamental de los tres campos, teoría de los test, escalamiento psicológico y psicofísico, algo que nadie volvería a intentar, salvo la reedición de 1954 de su propio libro (Guilford, 1954), pues se habían vuelto lo suficientemente complejos como para exigir cada uno de ellos un tratamiento aparte. Esos años de incubación y desarrollo dan lugar también a la institucionalización, y así, en 1936, se funda la sociedad psicométrica americana con Thurstone a la cabeza y cuyo órgano de expresión será la revista *Psychometrika.* Como complemento de carácter aplicado, frente al más teórico de *Psychometrika,* aparecerá unos años más tarde *Educational and Psychological Measurement* (1941); luego seguirían otras muchas entre las que cabe destacar *The British Journal of Mathematical and Statistical Psychology* (1948), *Journal of Educational Measurement* (1964), *Journal of Educational Statistics* (1976) y *Applied Psychological Measurement* (1977), por citar algunas.

En 1947 Thurstone publica su clásico texto *Análisis factorial múltiple,* técnica estadística con orígenes en el campo psicométrico (Burt, 1941; Kelley, 1928; Spearman, 1927), que aportará un notable avance para la construcción, análisis y validación de los test. En un campo como el del análisis factorial y las técnicas multivariadas en general, en el que se han producido avances tan gigantescos, potenciados por las facilidades informáticas y los refinamientos estadísticos actuales, resulta, sin embargo, refrescante releer el libro de Thurstone y admirar la cordura psicológica que lo impregna. Y es que cuando se piensa que un análisis factorial que un ordenador personal actual despacha en unos segundos podía llevar meses a un equipo de investigadores de entonces, se entiende que se afinasen las hipótesis y aguzase el ingenio antes de someterlas a prueba.

Por los años cuarenta también publicará Stevens (Stevens, 1946) su famoso trabajo sobre las escalas de medida, que obligará a los estudiosos de la teoría de los test a plantearse el estatus teórico de sus mediciones, además de sus propiedades empíricas, así como a terciar en la polémica que se abre entonces, y llega hasta nuestros días (Borsboom, 2005; Gaito, 1980; Michell, 1986; Townsend y Ashby, 1984), sobre las relaciones entre las escalas y sus implicaciones en el uso de las distintas técnicas estadísticas.

Pero la síntesis por antonomasia de la teoría clásica de los test será realizada por Gulliksen (1950) en su más que clásico libro *Theory of Mental Test,* que constituye, sin duda, la exposición mejor estructurada del corpus clásico. Gulliksen, antiguo estudiante, luego ayudante y colega de Thurstone, reconoce explícitamente el papel del maestro en su formación, y en especial de su libro *Fiabilidad y validez de los test* (Thurstone, 1931), ya agotado por las fechas en las que Gulliksen escribía el suyo. También en los años cincuenta aparecerán las primeras recomendaciones técnicas para el uso de los test (*Technical Recommendations for Psychological Test and Diagnostic Techniques,* 1954), que conocerán diversas actualizaciones y revisiones, la última en 2014.

También el escalamiento tendrá su clásico en los años cincuenta con el libro de Torgerson (1958) *Theory and Methods of Scaling.* Las dos ramas hermanas, teoría de los test y escalamiento, seguirán sus propios caminos, y aunque en esta como en otras divisiones hay algo de arbitrario, pues la mayoría de los modelos podrían generalizarse tanto a personas como a estímulos (Mosier, 1940, 1941), también es verdad que existían bastantes problemas específicos que justificaban la división.

Florecer parece haber sido el sino de los años sesenta, y la teoría de los test no iba a ser menos. Precisamente, en 1968 aparecerá el libro de Lord y Novick (1968) *Statistical Theories of Mental Test Scores,* que sintetiza y reanaliza críticamente todo lo hecho anteriormente en teoría clásica de los test, abriendo nuevas y prometedoras perspectivas. En el libro se incluye además el trabajo de Birnbaum sobre los modelos de rasgo latente, que abrirán una línea completamente nueva en la teoría de los test, conocida hoy como «teoría de respuesta a los ítems» (TRI). Este vuelco del modo de hacer clásico propiciado por la TRI va a oscurecer la mejora, al menos teórica, que prometía aportar al planteamiento clásico la teoría de la generalizabilidad (TG) propuesta

por Cronbach y colaboradores (Cronbach, Rajaratnam y Gleser, 1963; Gleser, Cronbach y Rajaratnam, 1965; Cronbach, Gleser, Nanda y Rajaratnam, 1972). La TG puede considerarse una extensión del modelo clásico, más que una alternativa. Mediante el uso masivo del análisis de varianza, la TG pretenderá analizar las fuentes del error de medida de un modo sistemático y desglosado, frente al tratamiento globalizado de la teoría clásica. Pero las aparatosas complicaciones introducidas en relación con las ventajas prácticas aportadas, unidas a la aparición en la escena psicométrica del enfoque alternativo de los modelos de TRI, relegarán la TG a un papel secundario en el campo de la teoría de los test.

No sería exacto decir que en el libro de Lord y Novick aparecen las primeras aportaciones sobre la TRI (véase Muñiz y Hambleton, 1992, para un estudio detallado del origen y desarrollo de estos modelos), pero la exposición y sistematización de Birnbaum, unidas a la plataforma publicitaria que supuso el libro, resultaron fundamentales para la rápida expansión de los modelos. De hecho ya hacía ocho años que Rasch había publicado su libro (Rasch, 1960) sobre el modelo logístico de un parámetro.

Sea como fuere, el libro de Lord y Novick, de áspera lectura y notación infernal, marca un antes y un después en la teoría de los test: terminaba una época, la clásica, y comenzaba otra nueva marcada por el predominio absoluto de la TRI. La nueva no negaba la anterior, aunque, como señalara Lord (1980), utilizará poco de ella para su formulación. Veamos a grandes rasgos cuáles eran las innovaciones que aportaba la TRI. La teoría clásica hallábase enfrentada con dos problemas de fondo importantes que no encontraban una solución satisfactoria en el marco clásico. Por un lado, la medición de las variables no era independiente del instrumento utilizado, algo así como si la longitud de los objetos dependiese del tipo de regleta. Por ejemplo, si la inteligencia de dos personas se mide con test distintos, los resultados de las mediciones no están en la misma escala; luego estrictamente no sabríamos cuál de las dos es más inteligente. Naturalmente, se había desarrollado todo un conjunto de soluciones técnicas para paliar el problema y poder equiparar las puntuaciones, pero se carecía de una solución digna de una medición aspirante al adjetivo de científica. En pocas palabras, las mediciones no eran invariantes respecto de los instrumentos de medida. Por otro lado, las propiedades de los instrumentos (test, ítems) dependían del tipo de personas utilizadas para establecerlas, por lo que en puridad no eran propiedades de los instrumentos de medida, sino de la interacción de estos con los objetos medidos. Por ejemplo, un ítem resultaría fácil o difícil en función de la muestra de personas utilizada. En suma, los instrumentos de medida no eran invariantes respecto de las personas evaluadas. Pues bien, los modelos de TRI permitirán dar una solución adecuada a esos dos problemas de fondo, y además aportarán todo un conjunto de avances tecnológicos complementarios para la construcción y análisis de los test.

Una década de investigación intensa sobre los distintos aspectos de los nuevos modelos, tanto a nivel teórico como aplicado, permitirá a Lord (1980) sintetizar en un libro hoy clásico los avances acumulados. El libro abre la década de los ochenta, que conocerá una expansión inusitada de la literatura psicométrica bajo la óptica de la TRI y revitalizará áreas que se encontraban atascadas, tales como los bancos de ítems, el sesgo o los test referidos al criterio, por citar solo algunas. Como señala Anastasi (1988) en la sexta edición de su incombustible *Psychological Testing,* la década de los ochenta ha sido un período de avances inusuales en teoría de los test, tanto respecto al progreso tecnológico como a la sofisticación teórica o la mejora de la responsabilidad profesional. Esta década vendrá marcada por una hegemonía clara de los modelos de TRI, como puede comprobarse echando una ojeada a los congresos de las sociedades psicométricas europea y americana, o a las revistas, y por la aparición de monografías independientes para cada una de las áreas ahora vigorizadas por la TRI. No obstante, no debe sacar el lector la impresión de que esta hegemonía de la TRI supone la muerte del enfoque clásico, ni mucho menos, pues la parsimonia y sencillez del modelo lineal clásico lo hacen apropiado en numerosas situaciones en las que la maquinaria pesada de la TRI no puede maniobrar con eficacia.

A la vista de este panorama general de la teoría de los test esbozado en las líneas precedentes, se comprenderá la limitada dimensión de lo que se pretende presentar en las páginas que siguen, cuyo objetivo no es otro que ofrecer al lector de lengua hispana una exposición clara y comprensiva de los aspectos fundamentales del enfoque clásico de los test y de la teoría de respuesta a los ítems, indican-

do en cada caso las fuentes adecuadas que permitan a quien lo desee una mayor profundización y preparación para comprender cabalmente los recientes avances de la teoría de los test, que no son pocos.

Nota histórica sobre los test

El desarrollo de la teoría de los test esbozado en el apartado anterior corre parejo con la evolución de los test concretos que van surgiendo, y como es natural, ambos desarrollos influyen el uno sobre el otro: los avances teóricos sobre los test y estos y sus problemas sobre los progresos teóricos. Bien es verdad que el avance no ha sido completamente sincronizado y en algunos momentos históricos una línea se ha adelantado por un tiempo a la otra, para luego volver a equilibrarse, como buenos vasos comunicantes que son. Así como el hacer suele preceder al pensar, los test como instrumentos se han anticipado a su fundamentación teórica. Los orígenes remotos de los primeros test podrían rastrearse según Du Bois (1970) allá por el año 3000 a.C., cuando los emperadores chinos evaluaban la competencia profesional de sus oficiales. Pero los orígenes más cercanos que darán lugar a los actuales test hay que ubicarlos en aquellas primeras pruebas sensomotoras utilizadas por Galton (1822-1911) en su famoso laboratorio antropométrico de Kensington. En 1884, durante la Exhibición Internacional sobre la Salud, que tuvo lugar en Londres, por la módica suma de tres peniques Galton medía a los visitantes todo un conjunto de índices antropométricos y sensomotores luego utilizados en sus estudios, como, por ejemplo, las dimensiones de la cabeza, estatura, longitud de los brazos extendidos, peso, fuerza de ambas manos, capacidad respiratoria, agudeza visual de ambos ojos, altura sentado, longitud del brazo, agudeza auditiva, tiempo de reacción visual y auditivo, precisión al dividir una línea en dos y tres partes iguales, error al estimar la apertura de un ángulo de 90 grados y otro de 60, etc. Hoy encontramos natural que al tratar de relacionar estas medidas con el funcionamiento intelectual no se encontrase conexión alguna, pero la hipótesis galtoniana de origen (Galton, 1883) tenía su lógica: si los datos sobre los que operamos han de ser filtrados por los sentidos, aquellos que contasen con sensores más finos dispondrían de un campo más amplio de actuación. También cabe a Galton el honor de ser el primero que aplicó la tecnología estadística para analizar los datos provenientes de sus test, labor que continuará Pearson. Como bien señala Boring (1950), si la década de los ochenta del siglo pasado viene marcada por Galton, la de los noventa vendrá por Cattell y la primera de este siglo por Binet. James McKeen Cattell (1860-1944) será el primero en utilizar el término «test mental» en su artículo «Mental test and measurements» publicado en la revista *Mind* en 1890, pero sus test, al igual que los de Galton, a quien por cierto admiraba, eran de carácter sensorial y motor fundamentalmente, y el análisis de los datos dejó clara la nula correlación entre este tipo de pruebas y el nivel intelectual de las personas (Wissler, 1901).

Será Binet (Binet y Simon, 1905a) quien dé un giro radical en la filosofía de los test al introducir en su escala tareas de carácter más cognoscitivo, encaminadas a evaluar aspectos como el juicio, la comprensión y el razonamiento, que según él constituían los componentes fundamentales del comportamiento inteligente. La puntuación de los niños en la escala de Binet y Simon se expresaba en términos de edad mental, que no era otra cosa que la edad cronológica de los niños que obtenían la misma puntuación media que el niño evaluado. Seguramente no es ajena al éxito alcanzado por la escala esta forma tan sencilla y comprensible para los no profesionales de expresar las puntuaciones de los niños. En la revisión de la escala que llevó a cabo Terman en la Universidad de Stanford, y que se conoce como la revisión Stanford-Binet (Terman, 1916), se utilizó por primera vez el cociente intelectual (CI) para expresar las puntuaciones. La idea era originaria de Stern, que en 1911 propuso dividir la edad mental (EM) entre la cronológica (EC), multiplicando por 100 para evitar los decimales:

$$CI = \frac{EM}{EC} \times 100$$

Esta fructífera veta de las escalas individuales de inteligencia abierta por Binet en 1905 y que se continúa hasta nuestros días había intuido, mejor que la hipótesis galtoniana, que si se desea evaluar el potencial intelectual hay que utilizar tareas cognoscitivamente complejas, que se asemejen de algún modo al tipo de cometidos intelectuales que se

pretende predecir. Puede decirse que el éxito de estas escalas para predecir el rendimiento escolar débese en gran medida al parecido de las tareas exigidas por ambos lados, escala y escuela. De hecho la causa próxima para que Binet pusiese manos a la obra de confeccionar su escala fue un encargo del Ministerio de Instrucción Pública para la detección y educación de los niños con deficiencias de inteligencia que asistían a las escuelas de París. Para una buena exposición de la escala, así como las sucesivas revisiones llevadas a cabo, véase Anastasi (1988).

El paso siguiente en el devenir histórico de los test vendrá marcado por la aparición de los test de inteligencia colectivos, propiciados por la necesidad del ejército estadounidense en 1917 de seleccionar y clasificar a los soldados que iban a tomar parte en la Primera Guerra Mundial. Un comité dirigido por Yerkes diseñó a partir de diverso material ya existente, especialmente de test inéditos de Otis, los hoy famosos test Alfa y Beta, el primero para la población general y el segundo para utilizar con analfabetos o reclutas sin dominio del inglés. Tras numerosas revisiones, estos test todavía siguen en uso. Debido a su éxito en el ejército, una vez finalizada la guerra la industria y el resto de las instituciones en general adoptaron en masa el uso de los test, conociéndose una expansión sin precedentes durante la próxima década, aunque no siempre en las mejores condiciones, debido por un lado a las limitaciones de los propios test, todavía un tanto rudimentarios, y por otro al uso de las puntuaciones más allá de lo que era razonable.

Con la experiencia acumulada iba quedando cada vez más claro que una puntuación global de inteligencia tal como la medían estos test no describía con suficiente precisión los diferentes aspectos del funcionamiento intelectual de las personas, y se imponía la evaluación de características más específicas cara a realizar pronósticos particulares más precisos. Si bien ello ya se venía haciendo de un modo más o menos sistemático, con la aparición de las hoy clásicas baterías de aptitudes habrá de esperar a que la técnica del análisis factorial dé sus frutos a partir de los años treinta y sobre todo cuarenta. Su producto más genuino serán las *aptitudes mentales primarias* de Thurstone (Thurstone, 1938; Thurstone y Thurstone, 1941), que conformaban lo que entonces se consideraban los componentes fundamentales del funcionamiento inteligente: comprensión verbal, fluidez verbal, aptitud numérica, aptitud espacial, memoria, rapidez perceptiva y razonamiento general. Dando más importancia a un factor general de inteligencia que articularía jerárquicamente otros factores de grupo (escuela inglesa), o reclamando un plano de igualdad para factores múltiples (escuela americana), el hecho central fue que el análisis factorial había permitido estructurar, no sin polémicas, la otrora genérica puntuación global de la inteligencia. En ocasiones los modelos alcanzaron grados de fragmentación rayanos en el desmenuzamiento, como en el caso de Guilford (1967), que propone nada menos que 120 rasgos intelectuales. Los distintos modelos darán lugar a numerosas baterías de test (PMA, DAT, GATB, TEA, etc.) de uso habitual actualmente. Excede por completo del cometido de estas líneas tratar de decir algo acerca del problema subyacente de la naturaleza de la inteligencia; la literatura al respecto es ciertamente abundante, y se aconseja al lector interesado la consulta de textos en español como los de Juan-Espinosa (1997) o Colom (1995, 2002).

A la vez que se producían los desarrollos citados en el campo de los test cognoscitivos, también los test de personalidad se beneficiaban de los avances técnicos que se iban produciendo, especialmente los derivados del análisis factorial y otras técnicas multivariadas afines. Suele citarse como origen próximo de los cuestionarios de personalidad de carácter psicométrico la hoja de datos personales utilizada por Woodworth en 1917 durante la Primera Guerra Mundial para la detección de neuróticos graves. En la actualidad, la sofisticación técnica en la construcción y análisis de los test de personalidad, que son legión (CEP, EPI, MMPI, 16PF, CPI, etc.), no se diferencia en nada de la utilizada con los test de aptitudes, si bien existen problemas específicos en unos y otros. Por su parte, el psiquiatra suizo Rorschach propone en 1921 su famoso test proyectivo de manchas de tinta, al que seguirán otros test proyectivos de muy distinto tipo de estímulos y tareas, aunque basados en la discutible asunción de la proyección, entre los que cabe citar el TAT, CAT, test de frustración de Rosenzweig, etc. Sin embargo, la técnica proyectiva que puede considerarse pionera es la asociación de palabras o test de asociación libre, descrita por Galton y utilizada incluso en el laboratorio por Wundt

y Cattell, aunque siguiendo las prácticas de Kraepelin y Jung pronto se asoció con la clínica, especialmente con la psicoanalítica.

Quede ahí esta somera nota histórica para ubicar al lector en el devenir de los test. Una excelente panorámica, así como la descripción y clasificación de los test más importantes, pueden verse en Anastasi y Urbina (1997). Para estudios históricos más detallados véanse Goodenough (1949), Du Bois (1970) o Geisinger y Usher-Tate (2016), y para una información enciclopédica y detallada sobre test concretos, véanse las sucesivas ediciones de los *Mental Mesurement Yearbooks* editados por Buros.

2. MODELO LINEAL CLÁSICO

Como ya se ha señalado en el apartado anterior, el modelo lineal clásico hunde sus raíces en los trabajos pioneros de Spearman (1904, 1907, 1913). En los años siguientes conoce un rápido desarrollo y Gulliksen (1950) recoge y sistematiza lo hecho hasta entonces. Lord y Novick (1968) llevarán a cabo una reformulación y análisis del modelo, a la vez que abren nuevas y fructíferas vías para la teoría de los test. Exposiciones menos técnicas del corpus clásico pueden consultarse en Guilford (1936, 1954), Magnuson (1967), Allen y Yen (1979), Thorndike (1982), Crocker y Algina (1986), y en español Yela (1987), Muñiz (2002), Martínez Arias et al. (2006) o Abad et al. (2011).

A continuación se exponen los aspectos fundamentales del modelo lineal clásico, su formulación, asunciones y deducciones más significativas. Aunque resulte obvio, no conviene perder de vista que el objetivo central del modelo es la estimación de los posibles errores cometidos cuando se utilizan los test para medir variables psicológicas. La necesidad de un modelo para estimar los errores de medida, en psicología como en cualquier otra ciencia, proviene del hecho elemental de que los errores no son observables directamente; lo que se obtiene directamente tras utilizar un instrumento de medida es el valor empírico «mezclado» con el error cuya cuantía se desea estimar. La lógica general para la estimación de los errores es común a todas las ciencias, pero la naturaleza de lo psicológico añadirá incluso algunas complicaciones adicionales.

TABLA 1.1
Formulación del modelo

Modelo:	$X = V + e$
Supuestos:	1. $V = E(X)$
	2. $\rho(v,e) = 0$
	3. $\rho(e_j,e_k) = 0$
Definición:	Dos test, j y k, se denominan paralelos si la varianza de los errores es la misma en ambos $[\sigma^2(e_j) = \sigma^2(e_k)]$ y también lo son las puntuaciones verdaderas de los sujetos ($V_j = V_k$).

El modelo establece que la puntuación empírica (X) que obtiene una persona en un test es igual a la suma de dos componentes: la puntuación verdadera (V) de la persona en ese test y el error de medida (e) cometido en la medición. Es razonable pensar que la puntuación empírica que obtiene una persona cuando se le aplica un test en un momento dado no coincida exactamente con su verdadera puntuación en ese test, pues en ese momento puntual la persona está afectada por múltiples factores de difícil control que inciden en su conducta. Una clasificación bastante exhaustiva de estas hipotéticas fuentes de error puede consultarse en Stanley (1971). Si estos factores perjudican a la persona, obtendrá una puntuación empírica más baja de la que verdaderamente le correspondería; si le benefician, la obtendrá superior. Ahora bien, cuando se pasa un test a una persona no hay manera de saber su puntuación verdadera, lo único que tenemos es su puntuación empírica, es decir, los puntos que saca en el test; su puntuación verdadera habrá que estimarla basándonos en los supuestos del modelo.

Supuesto 1

La puntuación verdadera (V) es la esperanza matemática de la empírica: $V = E(X)$, donde X es la variable aleatoria «puntuación empírica de la persona». Este primer supuesto constituye en realidad una definición de la puntuación verdadera. El lector poco familiarizado con el concepto de esperanza

matemática puede hacerse una idea imaginando que se aplicase un test «infinitas» veces a la misma persona. También debe imaginar que cada aplicación no afecta a las otras y que la persona no cambia en el curso de las aplicaciones. En estas condiciones, la puntuación verdadera de la persona en el test sería la media aritmética de las puntuaciones empíricas obtenidas en las «infinitas» aplicaciones. La puntuación verdadera es, por tanto, un concepto matemático. A partir de los valores de X (puntuaciones empíricas), y bajo ciertos supuestos que se irán viendo, la TCT permite hacer estimaciones probabilísticas razonables acerca del valor de las puntuaciones verdaderas (V). Conviene entenderlo bien, pues a menudo se ha hecho una conceptualización platónica de las puntuaciones verdaderas, considerándolas algo mágico y estático, propiedad de las personas y que determina su conducta. Del modelo no se sigue esta interpretación circular; la puntuación empírica en un test es una muestra de conducta que si reúne ciertos requisitos de medida, y bajo ciertos supuestos, permite hacer inferencias probabilísticas fundadas. De esto trata la teoría de los test.

Supuesto 2

Se asume que no existe correlación entre las puntuaciones verdaderas de las personas en un test y sus respectivos errores de medida: $\rho(v, e) = 0$. En principio no hay razón para pensar que el tamaño de los errores vaya sistemáticamente asociado al tamaño de las puntuaciones verdaderas.

Supuesto 3

Los errores de medida de las personas en un test no correlacionan con sus errores de medida en otro test distinto: $\rho(e_j, e_k) = 0$. Si se aplican correctamente los test, los errores serán aleatorios en cada ocasión, y a priori no existe razón para que covaríen sistemáticamente unos con otros.

Hay que señalar que ninguna de las asunciones del modelo es comprobable empíricamente de un modo directo tal como están expresadas; por tanto, aunque plausibles y sensatas a priori, habrá que hacer deducciones que sí se puedan contrastar y confirmen o falseen el modelo. Novick (1966) y Lord y Novick (1968) ofrecen formulaciones axiomáticas rigurosas del modelo a las que se remite al lector ávido de elegancia matemática.

Definición

Finalmente, se definen teóricamente los test paralelos, asumiéndose implícitamente que se pueden construir de hecho. De un modo menos formal que el señalado en su definición, podría decirse que dos test se consideran paralelos si miden lo mismo pero con diferentes ítems. Lord y Novick (1968) han desarrollado además otros tipos de paralelismo. Denominan test «tau equivalentes» a aquellos con puntuaciones verdaderas iguales para las personas en ambas formas, pero con varianzas error no necesariamente iguales. Test «esencialmente tau equivalentes» serían aquellos en los que la puntuación verdadera de cada persona en uno de ellos es igual a la del otro más una constante:

$$V_1 = V_2 + K$$

Los autores desarrollan las implicaciones y extensiones de estas redefiniciones de paralelismo para la TCT.

3. DEDUCCIONES INMEDIATAS DEL MODELO

A continuación se presentan algunas de las deducciones que se siguen directamente del modelo y que se utilizarán más adelante. Véase en el apéndice su obtención.

$$e = X - V \qquad [1.1]$$

El error de medida (e) se define, por tanto, como la diferencia entre la puntuación empírica (X) y la verdadera (V).

$$E(e) = 0 \qquad [1.2]$$

La esperanza matemática de los errores de medida es cero; luego son errores insesgados.

$$\mu_x = \mu_v \qquad [1.3]$$

La media de las puntuaciones empíricas es igual a la media de las verdaderas.

$$\text{cov}\,(V, e) = 0 \qquad [1.4]$$

Las puntuaciones verdaderas no covarían con los errores, lo cual es inmediato del supuesto 2 del modelo.

$$\text{cov}\,(X, V) = \text{var}\,(V) \qquad [1.5]$$

La covarianza entre las puntuaciones empíricas y las verdaderas es igual a la varianza de las verdaderas.

$$\text{cov}\,(X_j, X_k) = \text{cov}\,(V_j, V_k) \qquad [1.6]$$

La covarianza entre las puntuaciones empíricas de dos test es igual a la covarianza entre las verdaderas.

$$\text{var}\,(X) = \text{var}\,(V) + \text{var}\,(e) \qquad [1.7]$$

La varianza de las puntuaciones empíricas es igual a la varianza de las verdaderas más la de los errores.

$$\rho(X, e) = \sigma_e/\sigma_x \qquad [1.8]$$

La correlación entre las puntuaciones empíricas y los errores es igual al cociente entre la desviación típica de los errores y la de las empíricas.

$$\mu_1 = \mu_2 = \cdots = \mu_k \qquad [1.9]$$

$$\sigma^2(X_1) = \sigma^2(X_2) = \cdots = \sigma^2(X_k) \qquad [1.10]$$

$$\rho(X_1, X_2) = \rho(X_1, X_3) = \cdots = \rho(X_j, X_k) \qquad [1.11]$$

Es decir, para K test paralelos, las medias [1.9], las varianzas [1.10] y las intercorrelaciones entre ellos [1.11] son iguales.

Las ocho primeras deducciones son meras tautologías, no se pueden contrastar empíricamente, mientras que sí se pueden someter a prueba empírica las tres últimas. Nótese que el modelo y las deducciones están formulados para los valores paramétricos de la población. En términos generales —ya se irá matizando esta afirmación—, la teoría clásica de los test asume que los estadísticos obtenidos en muestras suficientemente amplias constituyen estimadores apropiados de los valores de la población. Cuanto más amplias sean las muestras, más pertinente será esta lógica. Lord y Novick (1968) proporcionan algunos estimadores más refinados para el caso de que las muestras no sean suficientemente amplias o no se cumplan estrictamente las condiciones de paralelismo.

2 Fiabilidad

Las mediciones psicológicas, como las de cualquier otra ciencia, han de ser fiables, es decir, han de estar libres de errores de medida. Un instrumento de medida, en nuestro caso un test o una escala, se considera fiable si las medidas que se hacen con él carecen de errores de medida, son consistentes. Una balanza es fiable si cada vez que pesamos el mismo objeto nos da el mismo resultado. Análogamente, un test será fiable si cada vez que se aplica a las mismas personas da el mismo resultado. La balanza lo tiene más fácil: el mismo objeto puede pesarse varias veces sin problema, pero los humanos cambian de una vez para otra, y a veces puede resultar problemático saber con seguridad si la inestabilidad observada en las mediciones se debe a la imprecisión del instrumento o a los cambios legítimos operados por las personas. Los errores de medida de los que se ocupa la fiabilidad son aquellos no sometidos a control e inevitables en todo proceso de medir, sea fisico, químico, biológico o psicológico. En muchas ocasiones las diferencias entre una medición y otra no dependen solo de estos errores, pudiendo explicarse además por los cambios operados en las personas, debidos a procesos madurativos, intervenciones o eventos de cualquier otro tipo. Incluso las inconsistencias pueden tener sentido en el marco en el que se lleva a cabo la medición. En estos casos la inestabilidad de las mediciones requiere una explicación y carece de sentido atribuirla a los errores aleatorios. La fiabilidad no trata ese tipo de «errores», que han de venir explicados por los modelos manejados. En cada situación el psicólogo tendrá que identificar las fuentes de error que afecten a las mediciones y no achacar, por ejemplo, a la baja fiabilidad de los instrumentos de medida lo que puede ser sencillamente variabilidad legítima de la variable medida. Nótese bien que el concepto de fiabilidad no se contradice en absoluto con la naturaleza cambiante de la conducta humana, como de un modo superficial se ha sugerido en algunas ocasiones; que cambie lo medido no anula la exigencia —todo lo contrario— de que los instrumentos de medida sean precisos. ¿Cómo medir adecuadamente el cambio sin instrumentos precisos? No conviene confundir la fiabilidad del instrumento de medida con la estabilidad o modificabilidad del constructo medido. La fiabilidad se refiere a la estabilidad de las mediciones cuando no existen razones teóricas ni empíricas para suponer que la variable a medir haya sido modificada diferencialmente para las personas, por lo que se asume su estabilidad, mientras no se demuestre lo contrario. Por ejemplo, parece sensato suponer que si se mide la inteligencia espacial de unas personas un día determinado y también se hace al día siguiente, su valor ha de ser básicamente el mismo, por lo que las posibles mínimas diferencias esperadas podrían atribuirse razonablemente a los errores aleatorios inherentes a todo acto de medir. Un test no sería fiable si cada día generase mediciones diversas de una variable que se supone estable. Ahora bien, lo que es válido para la inteligencia espacial no tiene por qué serlo para otras variables; por ejemplo, parece que la hora del día puede explicar gran parte de la variabilidad del tiempo de reacción de las personas, por lo que será en cada caso el psicólogo quien investigue las fuentes de error de las mediciones. No obstante, a nadie se le ocurrirá decir que los relojes que miden el tiempo de reacción en milisegundos no

son fiables por el hecho de que las medidas varíen a lo largo del día; la fiabilidad del instrumento no va unida a la estabilidad de la variable medida a lo largo del tiempo. Para un análisis detallado de la dialéctica fiabilidad/estabilidad de la conducta desde el punto de vista de la psicología clínica, véanse Silva (1989) y en general Cronbach y Furby (1970). Una interesante revisión sobre el problema de la fiabilidad en las investigaciones psicológicas puede verse en Schmidt y Hunter (1996).

En este capítulo se analizarán los distintos modos de estimar la fiabilidad de los test y la problemática implicada.

1. COEFICIENTE DE FIABILIDAD

El coeficiente de fiabilidad, $\rho_{XX'}$, se define como la correlación entre las puntuaciones obtenidas por las personas evaluadas en dos formas paralelas de un test, X y X'.

Es un indicador de la estabilidad de las medidas, pues si aplicamos un test X a una muestra de personas y pasado un tiempo aplicamos a las mismas personas una forma paralela X', dado que ambas formas miden lo mismo, si no hubiese errores aleatorios de medida, la correlación debería ser perfecta: $\rho_{XX'} = 1$. Por tanto, el grado en el que $\rho_{XX'}$ se aleja de 1 nos indicará en qué medida nuestras mediciones están afectadas por errores aleatorios de medida, siempre en el supuesto, claro está, de que las dos formas paralelas, X y X', realmente lo sean.

De la definición dada para el coeficiente de fiabilidad y de los supuestos del modelo se deriva fácilmente (véase apéndice):

$$\rho_{XX'} = \frac{\sigma_V^2}{\sigma_X^2} \qquad [2.1]$$

$$\rho_{XX'} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \qquad [2.2]$$

A partir de estas dos fórmulas es imposible calcular empíricamente $\rho_{XX'}$, dado que el valor de $(\sigma_v)^2$ y $(\sigma_e)^2$ no se puede obtener de las respuestas de las personas a los ítems. No obstante, son útiles para dar una idea conceptual de lo que representa el coeficiente de fiabilidad. Por la primera, [2.1], se ve que $\rho_{XX'}$ indica la proporción que la varianza verdadera es de la empírica. Si no hubiese errores aleatorios, entonces $(\sigma_v)^2 = (\sigma_x)^2$ y $\rho_{XX'} = 1$. Tal vez se vea más claro todavía en [2.2]: si $(\sigma_e)^2 = 0$, $\rho_{XX'} = 1$; y si $(\sigma_e)^2 = (\sigma_X)^2$, o, lo que es lo mismo, $(\sigma_v)^2 = 0$, entonces $\rho_{XX'} = 0$. Se suele denominar índice de fiabilidad (ρ_{xv}) a la correlación entre las puntuaciones empíricas de un test y las verdaderas, siendo igual a la raíz cuadrada del coeficiente de fiabilidad (véase apéndice):

$$\rho_{XV} = \sqrt{\rho_{XX'}} = \frac{\sigma_V}{\sigma_X} \qquad [2.3]$$

Error típico de medida

Se denomina error típico de medida (σ_e) a la desviación típica de los errores de medida (e). Despejando de [2.2], su fórmula viene dada por:

$$\sigma_e = \sigma_X \sqrt{1 - \rho_{XX'}} \qquad [2.4]$$

2. ESTIMACIÓN EMPÍRICA DEL COEFICIENTE DE FIABILIDAD

Como ya se ha señalado, las fórmulas del coeficiente de fiabilidad expuestas hasta ahora no permiten calcular su valor empírico para una muestra determinada de personas. Para poder hacerlo hay que valerse de su definición: correlación entre las puntuaciones en dos formas paralelas.

Se tratará, en suma, de:

1. Elaborar las dos formas paralelas.
2. Aplicarlas a una muestra amplia de personas representativas de la población en la que se va a utilizar el test.
3. Calcular la correlación entre las puntuaciones de las personas en ambas formas.

Dicha correlación será precisamente el coeficiente de fiabilidad. Este método se denomina por razones obvias *método de las formas paralelas,* y es el que emana genuina y directamente del modelo. No es infrecuente denominar coeficiente de equivalencia al valor obtenido, aludiendo a que cierta-

mente indicaría el grado en el que ambas formas son equivalentes. Se suelen utilizar además otros dos métodos, denominados, respectivamente, «test-retest» y «dos mitades». Veamos en qué consisten.

Test-retest. Para calcular el coeficiente de fiabilidad por este método se aplica el mismo test en dos ocasiones a las mismas personas; la correlación entre las puntuaciones de las dos aplicaciones será el coeficiente de fiabilidad. Dado que obviamente un test es paralelo a sí mismo, este método es perfectamente congruente con el modelo, denominándose a la estimación obtenida «coeficiente de estabilidad», pues indica en qué grado son estables las mediciones realizadas en la primera aplicación del test.

Dos mitades. Por este método se aplica el test una sola vez, obteniéndose para cada persona las puntuaciones correspondientes a cada una de las mitades en las que se divide el test. El coeficiente de fiabilidad viene dado por la correlación entre esas dos mitades (que será la estimación de la fiabilidad del test mitad) *más una corrección* para obtener la fiabilidad del test total (esta corrección se verá más adelante cuando se exponga la fórmula de Spearman-Brown). La estimación así obtenida, más que equivalencia o estabilidad, como en los casos anteriores, indica la covariación o consistencia interna de las dos mitades; es, pues, un indicador de la consistencia interna del test.

Si bien la lógica de estos tres métodos es clara, su realización empírica plantea diversos problemas experimentales relativos a la validez interna, para los cuales el modelo no da especificaciones concretas, quedando al criterio del psicólogo para cada situación nueva planteada. A continuación se comentan algunos de ellos.

En el método de las formas paralelas el problema fundamental es la construcción de dichas formas paralelas. Es difícil a nivel teórico hacer un test que mida exactamente lo mismo que otro, pero con distintos ítems; tal vez, incluso, filosóficamente imposible, y en la práctica es enormemente laborioso. Si se superan los problemas y se dispone de dos (o más) formas paralelas, probablemente es el método más recomendable.

En el método test-retest una cuestión de difícil solución es delimitar el tiempo óptimo que debe transcurrir entre ambas aplicaciones. Si se deja mucho, se introduce una gran fuente de invalidez interna, a saber, la ignota influencia diferencial de ese período de tiempo en las personas; pero si transcurre poco tiempo, la invalidez interna se cuela vía memoria de lo realizado previamente. No hay regla universal: depende en gran parte del tipo de test, ya que es evidente que hay unos test más propensos a ser recordados que otros. Una aproximación rigurosa al problema general del test-retest puede consultarse en Jöreskog y Sörbom (1976).

Finalmente, el método de las dos mitades es muy funcional, pues solo exige una sola aplicación del test. No obstante, hay que garantizar que las mitades del test sean paralelas. No es recomendable, por ejemplo, considerar mitades la primera parte del test por un lado y la segunda por otro, pues las personas evaluadas llegarán más cansadas a la segunda; además, en muchos test cognoscitivos los ítems van aumentando en dificultad, por lo que la segunda parte resultaría más difícil que la primera. Para evitar esto es frecuente tomar como una mitad los ítems pares y como otra los impares, o usar algún otro tipo de apareamiento de los ítems. En definitiva, es un problema de control experimental.

Un factor del test a tener en cuenta para elegir un método u otro de los comentados, o de otros que se verán más adelante, es si se trata de un test de velocidad o de un test de potencia. Suele entenderse por test de velocidad aquel cuya realización no conlleva dificultad alguna, o, más exactamente, el que todas las personas son capaces de realizar, aunque difieran en la rapidez de ejecución. Por el contrario, un test de potencia o poder sería aquel en el que las diferencias entre las personas son generadas por su distinta capacidad intelectual para resolver las tareas propuestas. No hacen falta muchas explicaciones para entender que en la práctica la mayoría de los test suelen ser mixtos, variando la proporción de ambos componentes: en unos predomina más la velocidad, y en otros, la potencia.

Índices de velocidad-potencia

El grado de velocidad de un test influye en sus parámetros más importantes, como su fiabilidad, su validez, la estructura factorial de los ítems, o en su caso de la batería; de ahí que se hayan propuesto

diferentes índices para expresar la proporción velocidad/potencia (véase, por ejemplo, Donlon, 1978, para un buen análisis). Gulliksen (1950) sugiere el cociente entre la varianza de los errores cometidos y la varianza de los fallos (errores más no intentados). Cuanto más bajo sea el cociente, más de velocidad será el test; con el límite cero indicando que todos los fallos se deben a no-intentos, el test sería de velocidad pura.

$$IV = \frac{\sigma_e^2}{\sigma_F^2}$$

Lord y Novick (1968) proponen un índice equivalente al de Cronbach y Warrington (1951) que indica la proporción de varianza atribuible a la velocidad:

$$IV = 1 - \frac{\rho_{VP}^2}{\rho_{VV'}\rho_{PP'}} \qquad [2.5a]$$

donde:

- IV: Índice de velocidad.
- ρ_{VP}: Correlación entre el test administrado en condiciones de velocidad (V), esto es, con tiempo limitado para su ejecución, y el test administrado con tiempo «ilimitado», en condiciones de potencia (P).
- $\rho_{VV'}$: Coeficiente de fiabilidad del test en condiciones de velocidad.
- $\rho_{PP'}$: Coeficiente de fiabilidad del test en condiciones de potencia.

El mayor inconveniente práctico de este índice es que requiere aplicar el test dos veces, lo que no siempre es fácil.

Ejemplo

El test de inteligencia general BLSIV se aplicó a una muestra de 500 universitarios con un tiempo de cinco minutos. A la semana siguiente se les aplicó sin tiempo límite. La correlación entre las puntuaciones de los universitarios en ambas aplicaciones fue 0,60; la fiabilidad del test con tiempo limitado, 0,70, y sin tiempo límite, 0,80. Calcular el índice de velocidad del test.

$$IV = 1 - \frac{(0{,}60)^2}{(0{,}70)(0{,}80)} = 0{,}36$$

El 36% de la varianza de las puntuaciones sería atribuible a la velocidad de respuesta de los universitarios.

Otro indicador más sencillo y de fácil uso es el cociente de velocidad propuesto por Stafford (1971):

$$CV = \frac{\sum NI}{\sum E + \sum O + \sum NI} \times 100 \qquad [2.5b]$$

donde:

- E: Número de errores de cada persona.
- O: Número de omisiones de cada persona.
- NI: Número de ítems no intentados por cada persona.

Cuando un test es de velocidad pura, ΣE y ΣO son cero, y en consecuencia $CV = 100$. En el caso de un test de potencia pura, $\Sigma NI = 0$, por lo que $CV = 0$, o, lo que es lo mismo, el cociente de potencia, que es el complementario del de velocidad, será 100.

Suprimiendo el sumatorio de la fórmula anterior, se tiene el cociente de velocidad para cada persona. Asimismo, si en vez de los datos correspondientes a una persona utilizamos los de un ítem, tendremos el cociente de velocidad del ítem.

3. ESTIMACIÓN DE LAS PUNTUACIONES VERDADERAS

Conocida la fiabilidad del test por alguno de los métodos expuestos, se pueden hacer ciertas estimaciones acerca de las puntuaciones verdaderas de las personas en el test, o, lo que es lo mismo, se pueden hacer estimaciones acerca de la cantidad de error que afecta a las puntuaciones empíricas. A continuación se exponen las tres estrategias más comúnmente utilizadas por la psicometría clásica, advir-

tiendo desde el principio que estas estimaciones han de tomarse con extremada cautela cuando se hacen para una persona en particular, siendo más apropiadas para la descripción de grupos.

a) *Estimación mediante la desigualdad Chebychev*

Como es sabido, la desigualdad de Chebychev establece que para toda variable X con media $\overline{X}$ y desviación típica S_X:

$$\forall K \quad P\{|X - \overline{X}| \leq K(S_X)\} \geq 1 - \frac{1}{K^2} \quad [2.6]$$

que traducido a la terminología psicométrica del modelo clásico:

$$\forall K \quad P\{|X - V| \leq K(\sigma_e)\} \geq 1 - \frac{1}{K^2} \quad [2.7]$$

Veamos el paso de la fórmula general [2.6] a la terminología psicométrica:

Se ha sustituido la media $\overline{X}$ por V, puesto que en el modelo clásico $E(X) = \mu_X = V$. Por su parte, la desviación típica S_X se ha sustituido por σ_e, puesto que en el modelo la varianza de X para un valor dado de V (persona o clase de personas) es igual a la varianza de los errores $(\sigma_e)^2$. Nótese que al fijar V lo que varían las empíricas se debe únicamente a la variación de los errores. Asimismo, la varianza de los errores para un determinado valor de V (persona o clase de personas) es igual a la varianza de los errores de la población σ_e^2 dado que $\rho_{ve} = 0$. En definitiva:

$$\sigma^2(X|V) = \sigma^2(e|V) = \sigma_e^2$$

Ejemplo

Se aplicó un test de rapidez perceptiva a una muestra representativa de 1.000 personas, obteniéndose una media de 70 puntos, una desviación típica de 10 y un coeficiente de fiabilidad de 0,64. Al nivel de confianza del 99%, ¿qué puntuación verdadera se estimará a las personas que obtuvieron una puntuación empírica en el test de 80 puntos?

Datos:

$N = 1.000;\ S_X = 10;\ r_{XX'} = 0{,}64;\ X = 80$

$$S_e = S_X\sqrt{1 - r_{XX'}} = 10\sqrt{1 - 0{,}64} = 6$$

$1 - 1/K^2 = 0{,}99$; por tanto: $K = 10$.
Sustituyendo en [2.7]:

$$P\{|80 - V| \leq (10)(6)\} \geq 0{,}99$$
$$P\{|V - 80| \leq (10)(6)\} \geq 0{,}99$$
$$P\{-60 \leq V - 80 \leq 60\} \geq 0{,}99$$
$$P\{20 \leq V \leq 140\} \geq 0{,}99$$

Al nivel de confianza del 99% estimamos que la puntuación verdadera en el test de rapidez perceptiva para las personas que obtuvieron una empírica de 80 estará ¡entre 20 y 140! Ciertamente es una pobre estimación; el intervalo es demasiado amplio, y ello se debe a dos razones: una, que el coeficiente de fiabilidad es bajo (0,64), y otra, que este método de estimación paga un caro tributo al no hacer ninguna asunción-restricción sobre la forma de la distribución de X, aunque, eso sí, es válido sea cual sea esta, lo cual es interesante, pero a un precio prohibitivo a base de abrir el intervalo confidencial.

b) *Estimación basada en la distribución normal de los errores*

Este es el método más utilizado en los textos clásicos. Una forma de evitar una estimación tan genérica como la anterior es asumir que los errores de medida, y por ende las puntuaciones empíricas, para un valor dado de V (persona o clase de personas con la misma puntuación verdadera) se distribuyen según la curva normal:

$$f(e|V) \sim N(0, \sigma_e^2)$$

Dado que $X = V + e$, con V constante, es inmediato que

$$f(X|V) \sim N(V, \sigma_e^2)$$

Adviértase que el precio a pagar por reducir la amplitud del intervalo es la asunción de normalidad que se añade al modelo. Hasta ahora no se había hecho ninguna asunción sobre la forma de las distribuciones de las puntuaciones. Esta asunción de normalidad e igualdad de las varianzas condicionales a lo largo de la escala de las puntuaciones (homoscedasticidad) ha sido cuestionada con frecuencia, especialmente en lo concerniente a los valores extremos de la escala. Si no se cumple, lo cual es bastante probable en la práctica, no sería muy preciso utilizar el mismo error típico de medida (ETM) para todas las personas, independientemente de su puntuación en el test, como se hace habitualmente. Según los datos de Feldt, Steffan y Gupta (1985), el valor máximo del ETM correspondería a la franja de puntuaciones medias, encontrando grandes diferencias para los distintos niveles de puntuaciones. Este problema no hallará una respuesta apropiada en el marco de la teoría clásica de los test, y habrá que esperar al desarrollo de los modelos de teoría de respuesta a los ítems para una solución adecuada. Mediante la función de información del test, estos modelos permitirán estimar el error de medida para los distintos niveles de competencia de las personas.

Supuesto esto, para estimar la puntuación verdadera en el test se establece el intervalo confidencial en torno a la puntuación empírica del modo habitual. Veámoslo para el ejemplo anterior y comparemos aquel intervalo con este:

1. Al NC del 99% corresponde una Z_C de ±2,58.
2. Error máximo admisible: $(Z_C)(\sigma_e) = = (2{,}58)\,(6) = 15{,}48$.
3. Intervalo confidencial:

$$(80 - 15{,}48) \leqslant V \leqslant (80 + 15{,}48)$$
$$(64{,}52 \leqslant V \leqslant 95{,}48)$$

Obsérvese la reducción del intervalo confidencial respecto al método anterior para los mismos datos, aunque sigue siendo excesivamente amplio debido a la baja fiabilidad del test.

c) Estimación según el modelo de regresión

Como es bien sabido, en el modelo de regresión lineal el pronóstico de una variable Y a partir de otra X, según el criterio de mínimos cuadrados, viene dado por la expresión:

$$Y' = \rho_{XY}\left(\frac{\sigma_Y}{\sigma_X}\right)(X - \bar{X}) + \bar{Y} \qquad [2.8]$$

Dado que ese es nuestro problema, estimar V a partir de X, traduzcamos esta expresión a nuestra terminología, donde Y, lo que se desea pronosticar, pasa a ser V, y X sigue siendo X:

$$V' = \rho_{XV}\left(\frac{\sigma_V}{\sigma_X}\right)(X - \bar{X}) + \bar{V}$$

Ahora bien, como hemos visto, $\bar{V} = \bar{X}$ y $\sigma_V/\sigma_X = \rho_{XV}$; luego:

$$V' = \rho^2_{XV}(X - \bar{X}) + \bar{X}$$

y como $\rho^2_{XV} = \rho_{XX'}$

$$V' = \rho_{XX'}(X - \bar{X}) + \bar{X} \qquad [2.9]$$

Mediante esta fórmula podemos hacer estimaciones puntuales de V a partir de X, conociendo el coeficiente de fiabilidad, la media del test y la puntuación empírica.

Ejemplo

Se aplicó un test de comprensión verbal a una muestra de 500 personas y se obtuvo una media de 40 y un coeficiente de fiabilidad de 0,80. ¿Qué puntuación verdadera en el test se pronosticará a las personas que obtuvieron una puntuación empírica de 60 puntos?

Sustituyendo en [2.9]:

$$V' = 0{,}80(60 - 40) + 40 = 56$$

Es decir, a las personas con una puntuación empírica de 60 se les pronostica una verdadera de 56. Ahora bien, el modelo de regresión utilizado lo único que garantiza es que «a la larga» los errores de

pronóstico cometidos son mínimos, según el criterio de mínimos cuadrados, pero la puntuación pronosticada V' no siempre coincidirá con V, denominándose precisamente a esa diferencia error de estimación. Es por ello por lo que para asegurarse de los pronósticos, en vez de realizar estimaciones puntuales, se establecen intervalos confidenciales en torno a la puntuación pronosticada V'. Para establecer dichos intervalos nos valemos del error típico de estimación, que es la desviación típica de los errores de estimación, y que en su forma general, para dos variables X e Y, viene dado por

$$\sigma_{Y \cdot X} = \sigma_Y \sqrt{1 - \rho_{XY}^2} \qquad [2.10]$$

Que traducido a la terminología y supuestos del modelo lineal clásico (véase apéndice) puede expresarse así:

$$\sigma_{V \cdot X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}} \qquad [2.11]$$

o también, teniendo en cuenta que $\sigma_X \sqrt{1 - \rho_{XX'}} = \sigma_e$:

$$\sigma_{V \cdot X} = \sigma_e \sqrt{\rho_{XX'}} \qquad [2.12]$$

Asumiendo que los errores de estimación se distribuyen normalmente en torno a V', se pueden establecer los correspondientes intervalos confidenciales. Establezcamos dicho intervalo para el mismo ejemplo de los apartados anteriores *a*) y *b*) y comparemos los resultados.

Recuérdese que los datos eran: NC: 99 %; $N = 1.000$; $S_X = 10$; $r_{XX'} = 0{,}64$; $X = 80$; $\bar{X} = 70$.

1. Al NC del 99 % corresponde una Z_C de ±2,58.
2. $\sigma_{V \cdot X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}} =$
   $= 10\sqrt{1 - 0{,}64}\sqrt{0{,}64} = 4{,}8$.
3. Error máximo:
   $$(Z_C)(\sigma_{V \cdot X}) = (2{,}58)(4{,}8) = 12{,}384.$$
4. $V' = \rho_{XX'}(X - \bar{X}) + \bar{X} = 0{,}64(80 - 70) + 70 =$
   $= 76{,}4$.
5. Intervalo confidencial: V' ± error máximo:
   $$(64{,}016 \leqslant V \leqslant 88{,}784)$$

Nótese que el intervalo así obtenido siempre será menor o igual al obtenido en el apartado *b*). Allí se utilizaba σ_e para establecerlo, mientras que aquí se utiliza $\sigma_{V \cdot X}$, que es igual a $\sigma_e \sqrt{\rho_{XX'}}$. Cuando $\rho_{XX'}$ tomase su valor máximo de 1 (por otra parte inalcanzable empíricamente), entonces $\sigma_{V \cdot X'} = \sigma_e$, y los intervalos serían iguales.

Nota. Todas las fórmulas utilizadas en este apartado están expresadas en puntuaciones directas. A continuación se ofrecen para la escala diferencial y típica. Se propone como ejercicio al lector la obtención de estas sencillas transformaciones de escala.

Directas	$Y' = \rho_{XY}\left(\frac{\sigma_Y}{\sigma_X}\right)(X - \bar{X}) + \bar{Y}$
diferenciales	$y' = \rho_{XY}(\sigma_Y/\sigma_X)x$
típicas	$Z_{Y'} = \rho_{XY} Z_X$
Directas	$V' = \rho_{XX'}(X - \bar{X}) + \bar{X}$
diferenciales	$v' = \rho_{XX'}x$
típicas	$Z_{Y'} = \sqrt{\rho_{XX'}} Z_X$
Directas	$\sigma_{Y \cdot X} = \sigma_Y \sqrt{1 - \rho_{XY}^2}$
diferenciales	Igual que en directas
típicas	$\sigma_{Z_x \cdot Z_Y} = \sqrt{1 - \rho_{XY}^2}$
Directas	$\sigma_{Y \cdot X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}}$
diferenciales	Igual que en directas
típicas	$\sigma_{Z_v \cdot Z_X} = \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}}$

4. FIABILIDAD DE LAS DIFERENCIAS

Existen numerosas situaciones aplicadas en psicología y educación en las que interesa estudiar las diferencias existentes entre las puntuaciones de las personas en un test y sus puntuaciones en otro; piénsese, por ejemplo, en la evaluación de intervenciones, terapias, tratamientos, etc. Para poder interpretar las diferencias encontradas es imprescindible disponer

de alguna medida de su fiabilidad. Dos diferencias iguales pueden tener muy distinto valor científico para el psicólogo en función de su fiabilidad.

Para dos test X y Z la fiabilidad de las diferencias entre sus puntuaciones: $(X - Z) = d$, como fácilmente se puede derivar (véase apéndice), viene dada por:

$$\rho_{dd'} = \frac{\sigma_X^2 \rho_{XX'} + \sigma_Z^2 \rho_{ZZ'} - 2\sigma_X \sigma_Z \rho_{XZ}}{\sigma_X^2 + \sigma_Z^2 - 2\sigma_X \sigma_Z \rho_{XZ}} \quad [2.13]$$

donde:

$\rho_{dd'}$: Coeficiente de fiabilidad de las diferencias.
σ_X^2: Varianza de las puntuaciones del test X.
σ_Z^2: Varianza de las puntuaciones del test Z.
$\rho_{XX'}$: Coeficiente de fiabilidad del test X.
$\rho_{ZZ'}$: Coeficiente de fiabilidad del test Z.
$\rho_{XZ'}$: Correlación entre ambos test.

Si ambos test se expresan en la misma escala para una mejor interpretabilidad de las diferencias, clásicamente en la escala de típicas, aunque cualquier otra es posible, en cuyo caso sus varianzas serían iguales ($\sigma_X^2 = \sigma_Z^2$), la fórmula anterior se simplifica:

$$\rho_{dd'} = \frac{\rho_{XX'} + \rho_{ZZ'} - 2\rho_{XZ}}{2(1 - \rho_{XZ})} \quad [2.14]$$

El error típico de medida de las diferencias (σ_{ed}), análogamente a lo visto para un solo test, vendrá dado por:

$$\sigma_{ed} = \sigma_d \sqrt{1 - \rho_{dd'}} \quad [2.15]$$

donde:

σ_d: Desviación típica de las diferencias.
$\rho_{dd'}$: Coeficiente de fiabilidad de las diferencias.

Ejemplo

El coeficiente de fiabilidad de un test de inteligencia espacial fue 0,60, y el de otro test también de inteligencia espacial, 0,50. La correlación obtenida entre las puntuaciones de 1.000 personas en ambos test fue de 0,40. ¿Cuál será el coeficiente de fiabilidad de las diferencias entre las puntuaciones de las personas en ambos test?

$$r_{dd'} = \frac{0{,}60 + 0{,}50 - 2(0{,}40)}{2(1 - 0{,}40)} = 0{,}25$$

La fiabilidad es muy baja. En una situación de este tipo más vale abstenerse de emitir juicios acerca de las diferencias halladas, pues son poco fiables.

Este tipo de análisis de las diferencias entre las puntuaciones de dos test (o más) es muy usual en el análisis de perfiles psicológicos. Como es bien sabido, un perfil psicológico no es otra cosa que la representación, generalmente gráfica, de las puntuaciones de una persona o de un grupo en determinadas variables, expresadas todas ellas en la misma escala (misma media y misma desviación típica) para percatarse mejor del sentido de las diferencias entre las variables.

El coeficiente de fiabilidad de las diferencias es, asimismo, fundamental para la medida del cambio experimentado por las puntuaciones de las personas en alguna variable psicológica o educativa. Casos típicos se presentan cuando hay que evaluar el impacto de terapias clínicas, programas de aprendizaje o cualquier otro tipo de intervención.

Conviene señalar finalmente que, en contra de la extendida costumbre, no está justificado hacer comparaciones individuales entre las puntuaciones empíricas de dos personas en un test, de una persona en dos test, etc., basándose en los errores típicos de medida que se pueden derivar (véase apartado siguiente). Este tipo de inferencias, como bien señalan Lord y Novick (1968), son incorrectas, y solo quedarían justificadas si los pares de personas que se desea comparar se cogiesen al azar y sin referencia a su puntuación empírica; si así se hiciese, a la larga (estrictamente infinito) las inferencias tendrían sentido globalmente, pero nunca para dos personas específicas elegidas, porque interesa comparar precisamente sus puntuaciones empíricas.

5. TIPOS DE ERRORES DE MEDIDA

Hasta aquí se han definido dos tipos de errores: el error de medida y el error de estimación. Cabe citar además (Gulliksen, 1950; Lord y Novick, 1968) el

error de sustitución y el error de predicción. A continuación se exponen los cuatro junto con sus respectivas desviaciones típicas y se comentan brevemente.

1. Error de medida

$$e = X - V$$
$$\sigma_e = \sigma_X \sqrt{1 - \rho_{XX'}} \qquad [2.4]$$

2. Error de estimación

$$e = V - V'$$
$$\sigma_{V \cdot X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}} \qquad [2.11]$$

3. Error de sustitución

$$e = X_1 - X_2$$
$$\sigma_{e(s)} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{2} \qquad [2.16]$$

4. Error de predicción

$$e = X_1 - X_1'$$
$$\sigma_{e(P)} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{1 + \rho_{XX'}} \qquad [2.17]$$

El error de medida y el error de estimación son, respectivamente, como ya se ha visto, la diferencia entre la puntuación empírica y la verdadera $(X - V)$, y la diferencia entre la puntuación verdadera y la verdadera pronosticada $(V - V')$. Sus desviaciones típicas, también previamente definidas, se denominan «error típico de medida» (σ_e) y «error típico de estimación» $(\sigma_{V \cdot X})$, respectivamente.

El error de sustitución es la diferencia entre las puntuaciones en un test (X_1) y las obtenidas en otro paralelo (X_2); es, por tanto, el error de medida que se generaría al sustituir una medición por otra proveniente de un test paralelo. Su desviación típica dada por [2.16] es el error típico de las diferencias entre dos test paralelos.

El error de predicción es la diferencia entre las puntuaciones en un test (X_1) y las puntuaciones pronosticadas en ese test (X_1') a partir de una forma paralela X_2. Es el error que se cometería al utilizar, en vez de las mediciones de un test, aquellas pronosticadas en ese test a partir de una forma paralela. Es decir, X_1' son los pronósticos realizados mediante la recta de regresión de X_1 sobre X_2, que viene dada según el modelo general [2.8] adaptado a nuestra terminología por:

$$X_1' = \rho_{12} \frac{\sigma_1}{\sigma_2} (X_2 - \bar{X}_2) + \bar{X}_1 \qquad [2.18]$$

A modo de ejercicio, trate el lector de derivar los cuatro errores típicos partiendo de las varianzas de los correspondientes errores de medida. Trate asimismo de ordenarlos de menor a mayor.

6. FACTORES QUE AFECTAN A LA FIABILIDAD

6.1. Fiabilidad y variabilidad

La fiabilidad de un test no depende únicamente de sus características propias, sino también del tipo de muestra de personas utilizadas para calcularla, lo cual constituye una seria limitación para el modelo clásico, pues se está describiendo un instrumento de medida, como es el test, en función de los «objetos» medidos, las personas. Uno de los aspectos de la muestra que influye en la fiabilidad de su variabilidad, el coeficiente de fiabilidad, aumenta al aumentar la variabilidad de la muestra. La razón es bien simple: el coeficiente de fiabilidad se ha definido como la correlación entre dos formas paralelas de un test, y es bien sabido que la correlación viene afectada por la variabilidad del grupo, aumentando con esta. Por tanto, un dato imprescindible para la interpretación del coeficiente de fiabilidad es la variabilidad de la muestra en la que se calculó; en otras palabras, un test no tiene un coeficiente de fiabilidad fijo, este depende de la variabilidad de la muestra en la que se calcule.

En suma, al aumentar la variabilidad de la muestra, aumenta el valor del coeficiente de fiabilidad, proponiéndose en este apartado una fórmula que permite estimar ese aumento. La fórmula es apropiada si se cumple el supuesto en el que se basa, a saber, que la varianza de los errores de medida en el test es la misma en ambas poblaciones, la menos variable y la más variable.

Según [2.4] el error típico de medida venía dado por:

$$\sigma_e = \sigma_X \sqrt{1 - \rho_{XX'}}$$

El supuesto citado puede expresarse, por tanto: $(\sigma_{e1})^2 = (\sigma_{e2})^2$, o, más explícitamente, sustituyendo $(\sigma_e)^2$ por su valor en ambas poblaciones:

$$\sigma_1^2(1 - \rho_{11'}) = \sigma_2^2(1 - \rho_{22'})$$

Despejando $\sigma_{22'}$:

$$\rho_{22'} = 1 - \frac{\sigma_1^2}{\sigma_2^2}(1 - \rho_{11'}) \qquad [2.19]$$

donde:

$\rho_{11'}$: Coeficiente de fiabilidad en la población 1.
$\rho_{22'}$: Coeficiente de fiabilidad en la población 2.
σ_1^2: Varianza empírica en la población 1.
σ_2^2: Varianza empírica en la población 2.

Ejemplo

El coeficiente de fiabilidad de un test de coordinación visomotora en una muestra de universitarios que habían superado las pruebas de selectividad fue de 0,64, obteniéndose una varianza de 25. ¿Cuál habría sido el coeficiente de fiabilidad del test si se hubiese calculado con los datos de todos los aspirantes cuya varianza fue 100, y no solo utilizando los que superaron la selectividad?

$$\rho_{22'} = 1 - \frac{25}{100}(1 - 0{,}64) = 0{,}91$$

Queda patente con el ejemplo la relatividad del coeficiente de fiabilidad, que pasa de 0,64 a 0,91 al aumentar la variabilidad de 25 a 100.

Adviértase que la pertinencia de la fórmula [2.19] puede comprobarse empíricamente aplicando el test a las dos poblaciones de interés y luego comparando los resultados hallados con los estimados por la fórmula. Como Lord y Novick (1968) señalaran, esta fórmula debe usarse con precaución, pues el fuerte supuesto en el que se basa no siempre se cumple. Si, por ejemplo, ocurriese que σ_{e1} fuese menor que σ_{e2} y utilizásemos la fórmula, se sobreestimaría $\rho_{22'}$, ya que, si $\sigma_{e1} < \sigma_{e2}$, entonces

$$\sigma_1\sqrt{1 - \rho_{11'}} < \sigma_2\sqrt{1 - \rho_{22'}}$$

con lo que

$$\rho_{22'} < 1 - \frac{\sigma_1^2}{\sigma_2^2}(1 - \rho_{11'})$$

6.2. Fiabilidad y longitud

La fiabilidad de un test también depende de su longitud, entendiendo por longitud el número de ítems que contiene. En principio parece enjundioso pensar que cuantos más ítems se utilicen para evaluar una variable, mejor podremos muestrear los diferentes aspectos que la conforman y más fiable será la medida obtenida. En el límite, infinitos ítems, el error sería cero; claro que la persona también sería cero: habría fenecido de una sobredosis de ítems; en matemáticas no, pero en psicología casi todo es cero en el límite.

Según los supuestos del modelo, si se tiene un test X y se aumenta su longitud n veces a base de ítems paralelos a los originales, la fiabilidad del nuevo test alargado viene dada por la conocida fórmula de Spearman-Brown, denominada con frecuencia «profecía de Spearman-Brown», dado que predice, profetiza, lo que le va a ocurrir a la fiabilidad del test al aumentar su longitud:

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1 + (n - 1)\rho_{xx'}} \qquad [2.20]$$

donde:

$\rho_{XX'}$: Fiabilidad del test alargado.
$\rho_{xx'}$: Fiabilidad del test original.
n: Número de veces que se ha alargado el test.

Un caso particular de esta fórmula especialmente popular es cuando se duplica la longitud del test original, utilizado, por ejemplo, en el cálculo del coeficiente de fiabilidad por el método de las dos mitades. En dicho caso particular $n = 2$, quedando [2.20] reducida a:

$$\rho_{XX'} = \frac{2\rho_{XX'}}{1 + \rho_{XX'}} \qquad [2.21]$$

Ejemplo

Un test que consta de 20 ítems se aplicó a una muestra de personas, obteniéndose un coeficiente de fiabilidad de 0,60. ¿Qué fiabilidad tendría el test si se le añadiesen 15 ítems paralelos a los que ya poseía?

$$n = \frac{20 + 15}{20} = 1{,}75$$

$$\rho_{XX'} = \frac{(1{,}75)(0{,}60)}{1 + (1{,}75 - 1)(0{,}60)} = 0{,}724$$

La nueva fiabilidad del test alargado sería 0,724, frente al 0,60 del test original.

Todo lo dicho respecto a alargar el test puede aplicarse a acortar, en cuyo caso n será menor que 1.

Ejemplo

Un test que consta de 150 ítems tiene un coeficiente de fiabilidad de 0,90. Dado que resulta tediosamente largo, se le han suprimido 60 ítems. ¿Cuál será la fiabilidad del nuevo test acortado?

$$n = \frac{150 - 60}{150} = 0{,}60$$

$$\rho_{XX'} = \frac{(0{,}60)(0{,}90)}{1 + (0{,}60 - 1)(0{,}90)} = 0{,}844$$

Al suprimir los 60 ítems, la fiabilidad baja de 0,90 a 0,844.

De la fórmula general [2.20], se puede despejar n, lo que permite estimar cuánto habría que alargar (o acortar) un test para obtener una fiabilidad determinada:

$$n = \frac{\rho_{XX'}(1 - \rho_{xx'})}{\rho_{xx'}(1 - \rho_{XX'})} \qquad [2.22]$$

Ejemplo

La fiabilidad de un test de 40 ítems resultó ser 0,80. ¿Cuántos ítems habría que añadirle para que su fiabilidad fuese 0,90?

$$n = \frac{0{,}90(1 - 0{,}80)}{0{,}80(1 - 0{,}90)} = 2{,}25$$

Hay que alargar el test 2,25 veces; luego el nuevo test alargado tendrá (2,25)(40) = 90 ítems; habría que añadirle 50 ítems (90 – 40 = 50).

Obtención de la fórmula de Spearman-Brown

Como se ha podido comprobar a través de los ejemplos anteriores, el uso de la fórmula de Spearman-Brown es ciertamente sencillo. Un par de preguntas que surgen inmediatamente es: ¿funciona la fórmula?, ¿son correctas las «profecías» hechas a partir de ella? Nótese que los resultados proporcionados por esta fórmula son susceptibles de someterse a comprobación empírica, puede estimarse la fiabilidad mediante ella y luego calcular la fiabilidad empíricamente, contrastando los resultados. La abundante literatura al respecto parece indicar que la fórmula funciona bastante bien en general, siempre y cuando, claro está, los ítems añadidos sean paralelos a los previamente existentes.

Para la obtención de la fórmula de Spearman-Brown se parte del concepto general de coeficiente de fiabilidad proporcionado por el modelo lineal clásico, es decir, el cociente entre la varianza de las puntuaciones verdaderas y la de las empíricas. Ahora bien, si se alarga n veces el test, tanto la varianza de las verdaderas del nuevo test alargado como la de las empíricas se verán afectadas. En concreto (véase apéndice), la varianza verdadera original queda multiplicada por n^2, es decir, la varianza verdadera del test alargado n veces será: $n^2\sigma_v^2$. Por su parte, la varianza empírica del test alargado (apéndice) viene dada por

$$n\sigma_x^2[1 + (n - 1)\rho_{xx'}]$$

Por tanto, la fiabilidad del test alargado será el cociente entre ambas:

$$\rho_{XX'} = \frac{n^2\sigma_v^2}{n\sigma_x^2[1 + (n - 1)\rho_{xx'}]}$$

y simplificando

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1 + (n - 1)\rho_{xx'}}$$

que es precisamente la fórmula propuesta.

Nótese que al aumentar n veces la longitud de un test su varianza verdadera aumenta proporcionalmente más que su varianza empírica, pues, mientras que la verdadera original resulta multiplicada por n^2, la empírica se multiplica por $n[1 + (n - 1)\rho_{xx'}]$, expresión cuyo valor solo se igualaría a n^2 en el caso de que $\rho_{xx'} = 1$, hecho poco probable en la práctica, por no decir imposible. Esta es precisamente la razón de que al aumentar la longitud aumente la fiabilidad, pues aumenta más el numerador (varianza verdadera) que el denominador (empírica).

Límite de $\rho_{xx'}$ cuando n tiende a infinito

Según Spearman-Brown:

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1 + (n - 1)\rho_{xx'}}$$

bajando n al denominador

$$\rho_{XX'} = \frac{\rho_{xx'}}{[1 + (n - 1)\rho_{xx'}]/n}$$

y dividiendo cada sumando entre n y simplificando:

$$\rho_{XX'} = \frac{\rho_{xx'}}{(1/n) + (n - 1)\rho_{xx'}/n}$$

$$\rho_{XX'} = \frac{\rho_{xx'}}{(1/n) + \rho_{xx'} - (1/n)\rho_{xx'}}$$

Ahora bien, cuando n tiende a infinito: $1/n$ es cero; luego

$$\rho_{XX'} = \rho_{xx'}|\rho_{xx'} = 1$$

Es decir, la fiabilidad de un test tiende a uno a medida que se aumenta su longitud, alcanzando teóricamente ese valor para infinitos ítems. Se recomienda cierta precaución en el salto de las matemáticas a la psicología. Mejorar la fiabilidad de un test a base de aumentar su longitud puede ser útil y recomendable, por ejemplo, cuando el test original ya tiene una fiabilidad razonable con no muchos ítems, pero si el test ya tiene un número considerable de ítems y, sin embargo, es poco fiable, más que su longitud hay que ir pensando en cambiar los ítems, en construir otro.

6.3. Fiabilidad y nivel de las puntuaciones en el test

Hasta ahora se ha visto cómo se calculaba el coeficiente de fiabilidad y el error típico de medida para una muestra determinada, asumiendo implícitamente que sus valores eran comunes para todas las personas de la muestra, independientemente de sus puntuaciones en el test. Ahora bien, un test no siempre resulta igualmente preciso para todas las personas; su error típico de medida puede depender de la puntuación o nivel de las personas en el test. En este apartado se expone la forma de calcular el error típico de medida para distintos niveles de puntuaciones en el test. Pero antes de pasar a exponer el método de cálculo, veamos a qué se puede deber esta variación de los errores típicos de medida. La causa fundamental de que el error típico de medida no sea el mismo para cualquier nivel de puntuaciones radica en el tipo de ítems que componen el test. Por ejemplo, si ocurriese que la mayoría de los ítems fuesen de dificultad media, el test mediría con mayor precisión a las personas de nivel medio, es decir, los errores de medida tenderían a ser mayores para el caso de personas de alta y baja competencias en la variable medida. Por el contrario, si los ítems son en su mayoría de dificultad elevada, la prueba tenderá a dar mediciones más precisas para las personas de alto nivel, en detrimento de aquellas con puntuaciones medias o bajas. Casos extremos serían aquellos en los que todos los ítems fuesen tan difíciles, o tan fáciles, que no fuesen contestados, respectivamente, por casi nadie o por la mayoría. La variación del error típico tiende a agudizarse a medida que aumenta la amplitud del rango de las puntuaciones en la variable medida. Además de la naturaleza de los ítems, pueden existir otros factores de carácter secundario que contribuyan también a que el error típico de medida no afecte por igual a todas las personas de la muestra, tales como unas instrucciones inadecuadas que induzcan a las personas con poco nivel a contestar al azar ante cualquier pregunta que

desconozcan, u otros por el estilo derivados de una aplicación incorrecta de la prueba.

Ante esta situación no parece apropiado usar el mismo error típico de medida para todas las personas, por lo que se recurre a la utilización de distintos errores típicos de medida en función de la cuantía de las puntuaciones de las personas en el test. Dado que no se puede generalizar de unos test a otros, ni de unas muestras a otras, habrá que calcular empíricamente en cada caso los errores típicos de medida correspondientes.

La forma más clásica de llevar a cabo el cálculo de los errores típicos de medida para los distintos niveles de puntuaciones fue propuesta por Thorndike (1951). Consiste en dividir las puntuaciones en varios niveles o categorías y calcular el error típico de medida para cada una de ellas. No se puede hablar de un número idóneo de categorías, que dependerá en gran medida del número de personas de la muestra. Un número mínimo podrían ser tres, puntuaciones bajas, medias y altas, pero si se dispone de suficientes sujetos, puede incrementarse el número de niveles, explorando de ese modo los errores con mayor exhaustividad a lo largo del rango de la variable medida. Si se dispusiese de las suficientes personas, podrían incluso hacerse tantas categorías como puntuaciones posibles en el test. Estrictamente hablando, los niveles deberían establecerse a partir de las puntuaciones verdaderas, pero en la práctica solo se dispone de las empíricas.

Si se dispone de una sola aplicación del test, una vez establecidas las categorías de las puntuaciones, se calcula el error típico de medida para cada una de ellas. Para ello se dividen las puntuaciones de cada persona en dos mitades (pares e impares, por ejemplo) y se calcula la desviación típica de las diferencias entre ambas. El resultado sería el error típico de medida para cada uno de los niveles de puntuaciones:

$$\sigma_e = \sigma_{(p-i)}$$

Nótese que si el test fuese perfectamente fiable, las puntuaciones de ambas mitades coincidirían, no habría errores de medida, el error típico de medida sería nulo y la fiabilidad perfecta. A medida que aumentan las diferencias entre las dos mitades, se incrementa el error típico de medida. Esta forma de conceptualizar el error típico de medida fue utilizado por Rulon (1939) para obtener su fórmula del coeficiente de fiabilidad (véase la fórmula 2.29). Si se denomina e al error del test global y e_1 y e_2 a los errores de cada una de las mitades, bajo los supuestos del modelo clásico, es fácil demostrar que la varianza de los errores globales del test es igual a la suma de las varianzas de los errores de cada una de las dos mitades:

$$\sigma_e^2 = \sigma^2(e_1 - e_1) = \sigma_{e1}^2 + \sigma_{e2}^2$$

Si se dispone de dos formas paralelas, o de dos aplicaciones del test, se forman las categorías a partir de la suma de las puntuaciones de cada persona en ambas formas y luego se procede al cálculo del error típico de medida para cada categoría. El error típico se obtiene calculando la desviación típica de las diferencias entre las dos formas del test y dividiendo el resultado por $\sqrt{2}$

$$\sigma_e = \frac{\sigma(X_1 - X_2)}{\sqrt{2}} \qquad [2.23]$$

En la fórmula 2.16 se expresa la desviación típica de las diferencias entre dos formas paralelas de un test; nótese cómo para obtener a partir de ella el error típico de medida hay que dividirla por $\sqrt{2}$.

A continuación se ilustran los cálculos anteriores mediante un ejemplo numérico.

Ejemplo

Se aplicó un test de aptitud espacial a una muestra de 12 personas. Las puntuaciones globales en el test, así como las obtenidas en las mitades par e impar, se ofrecen a continuación. Veamos cómo se calcula el error típico de medida para tres niveles de la variable medida: bajo, medio y alto.

	Personas	Global	Par	Impar	Diferencia ($P - I$)
Nivel bajo	A	6	4	2	2
	B	7	4	3	1
	C	9	5	4	1
	D	10	5	5	0

	Personas	Global	Par	Impar	Diferencia ($P - I$)
Nivel medio	E	12	6	6	0
	F	14	7	7	0
	G	15	8	7	1
	H	19	10	9	1
Nivel alto	I	22	12	10	2
	J	24	12	12	0
	K	26	14	12	2
	L	29	15	14	1

Una vez divididas las puntuaciones en los tres niveles exigidos, el error típico de medida de cada nivel será la desviación típica de las diferencias ($p - i$) de cada uno de los niveles. Aplicando la fórmula de la desviación típica a la columna de las diferencias, se obtienen los siguientes resultados para cada uno de los niveles:

— Bajo:

$$S_e = S_{(p-i)} = \sqrt{\frac{(2^2 + 1^2 + 1^2 + 0^2)}{4} - \left(\frac{4}{4}\right)^2} = 0{,}71$$

— Medio:

$$S_e = S_{(p-i)} = \sqrt{\frac{(0^2 + 0^2 + 1^2 + 1^2)}{4} - \left(\frac{2}{4}\right)^2} = 0{,}50$$

— Alto:

$$S_e = S_{(p-i)} = \sqrt{\frac{(2^2 + 0^2 + 2^2 + 1^2)}{4} - \left(\frac{5}{4}\right)^2} = 0{,}83$$

Como se puede observar, el error típico es menor para el nivel intermedio, lo cual quiere decir que el test está midiendo con más precisión a las personas con un nivel de tipo medio que a las de niveles bajo y alto. Si hubiese que estimar la puntuación verdadera de una persona en el test, sería aconsejable utilizar el error típico de medida correspondiente, en función del nivel en el que estuviese ubicada esa persona según su puntuación en el test. Compruebe el lector que el valor del error típico de la muestra global sin tener en cuenta los tres niveles vale 0,76, y se obtiene al calcular la desviación típica de la columna de las diferencias ($p - i$).

Se ha expuesto la forma de cálculo del error típico de medida, o fiabilidad absoluta, para los distintos niveles de la variable medida; a partir de esos valores se podría calcular también la fiabilidad relativa o coeficiente de fiabilidad para esos mismos niveles. No obstante, lo más habitual es el cálculo del error típico de medida, dado que va a permitir establecer intervalos confidenciales en torno a las puntuaciones empíricas para estimar las verdaderas. La gran ventaja de disponer de varios errores típicos de medida por intervalos es que va a permitir utilizar un error típico u otro en función del intervalo en el que se encuentre la puntuación a estimar. Por ejemplo, para los datos anteriores, si se desea estimar la puntuación verdadera de una persona cuya empírica en el test fue de 26 puntos, se utilizará el error típico correspondiente, es decir, 0,83. Por el contrario, si la puntuación empírica fuese 15, el error típico que habría que utilizar sería 0,50, etc.

A veces los test se utilizan para clasificar a los examinados en dos grupos, los que superan la prueba y los que no la superan, para lo cual hay que establecer un punto de corte. En estos casos es muy importante conocer cuál es el error típico en la zona en la que se produce el corte, pues los errores en esa zona tienen una gran incidencia sobre los fallos clasificatorios. La razón es sencilla: si la puntuación de una persona está cerca del punto de corte, cualquier mínimo error puede pasarla de un lado a otro del corte, mientras que si la puntuación está alejada del punto de corte, ese mismo error probablemente no tendrá influencia sobre la clasificación. Para conocer el error típico en esa zona vital se establece un intervalo en torno al punto de corte y mediante cualquiera de los procedimientos descritos se estima el error típico de medida. Con la amplitud que debe tener este intervalo pasa como con los niveles: no se puede establecer de una vez por todas, y dependerá del número de sujetos disponibles y del ajuste deseado en torno al punto de corte. Para una ampliación del estudio de la fiabilidad en relación con los puntos de corte, véase el apartado 2.9 sobre los test referidos al criterio.

Señalar finalmente que el método expuesto para el cálculo del error típico a distintos niveles es el más clásico, sugerido por Thorndike (1951), pero desde

entonces se han propuesto otros muchos. Por ejemplo, el lector interesado puede consultar el trabajo de Lord (1984), en el que el autor compara cuatro métodos distintos para estimar el error típico: los de Feldt, Steffan y Gupta (1985), donde se comparan cinco métodos, o el de Qualls (1992), donde son seis los métodos analizados. Feldt y Qualls (1996) han propuesto una variante interesante del método de Thorndike. En general, las diferencias de funcionamiento entre los distintos métodos no son notables, por lo que la elección de uno u otro, como bien señalan Feldt et al. (1985), va a depender de consideraciones prácticas y de las preferencias de los usuarios. Esta temática psicométrica relativa a la estimación de los errores típicos a distintos niveles va a dar un giro radical dentro del marco de la teoría de respuesta a los ítems, donde la precisión a lo largo de la escala de la variable medida vendrá dada por una curva denominada «función de información», que se expone en el apartado 7.7.

7. COEFICIENTE ALFA (α)

7.1. Concepto y fórmula

El coeficiente *alfa* (α), propuesto por Cronbach (1951), constituye otra forma de acercarse a la fiabilidad. Más que la estabilidad de las medidas, α refleja el grado en el que covarían los ítems que constituyen el test; *es, por tanto, un indicador de la consistencia interna del test.* Su fórmula viene dada por:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n}\sigma_J^2}{\sigma_X^2}\right) \qquad [2.24]$$

donde:

n: Número de ítems del test.
$\Sigma\sigma_J^2$: Suma de las varianzas de los n ítems.
σ_X^2: Varianza de las puntuaciones en el test.

Que α es función directa de las covarianzas entre los ítems, indicando, por tanto, la consistencia interna del test, tal vez se pueda apreciar más directamente si se expresa su fórmula en función explícita de dichas covarianzas (véase apéndice), que viene dada por:

$$\alpha = \frac{n}{n-1}\left(\frac{\sum_{j\neq k}^{n}\sum \mathrm{cov}(j,k)}{\sigma_X^2}\right) \qquad [2.25]$$

Es claro, según [2.25], que α aumenta al aumentar las covarianzas entre los ítems.

Conviene observar finalmente que, en contra de la idea tan extendida de usar el coeficiente α como un indicador preciso de la unidimensionalidad de los ítems, este tiene, sin embargo, algunas limitaciones al respecto. Es evidente que α va a resultar elevado si los ítems se acercan a la unidimensionalidad, pero lo contrario no es estrictamente cierto, y como señalan Green, Lissitz y Mulaik (1977), una elevada consistencia interna no implica necesariamente unidimensionalidad. Según los citados autores, α viene afectado por diversos factores para ser un índice apropiado de la unidimensionalidad:

1. Aumenta cuando se incrementa el número de ítems.
2. Aumenta cuando se repiten ítems similares.
3. Aumenta cuando el número de factores pertenecientes a cada ítem aumenta.
4. Fácilmente se acerca a 0,80 y lo supera cuando el número de factores que pertenecen a cada ítem es dos o más y el número de ítems moderadamente amplio.
5. Disminuye moderadamente al disminuir las comunalidades de los ítems.

Por todo ello, cuando se desee hacer juicios acerca de la unidimensionalidad, además de acerca de la consistencia interna, alfa debe complementarse con otras técnicas.

a) Estimador insesgado de α

Feldt, Woodruff y Salih (1987) han propuesto un estimador insesgado de α que viene dado por:

$$\bar{\alpha} = \frac{(N-3)\hat{\alpha}+2}{N-1} \qquad [2.26]$$

donde:

$\bar{\alpha}$: Estimador insesgado: $[E(\bar{\alpha}) = \alpha]$.
$\hat{\alpha}$: Valor de α obtenido en una muestra mediante [2.24].
N: Número de personas de la muestra.

A medida que aumenta el número de personas de la muestra (N), el valor hallado en la muestra y el estimador insesgado se acercan, siendo iguales cuando N tiende a infinito (véase apéndice).

Por ejemplo, para una muestra de 103 casos con un $\hat{\alpha} = 0{,}70$, el valor del estimador vendría dado por:

$$\bar{\alpha} = \frac{(103 - 3)(0{,}70) + 2}{103 - 1} = 0{,}706$$

una diferencia de 6 milésimas para un $N = 103$. A nivel práctico puede decirse que a partir de l00 sujetos las diferencias entre el valor sesgado y el insesgado son más bien irrelevantes.

b) α como límite inferior de $\rho_{XX'}$

Se demuestra (véase apéndice) que α es menor o igual que $\rho_{XX'}$, siendo igual cuando los ítems son paralelos, tau (τ) equivalentes o esencialmente tau (τ) equivalentes:

$$\alpha \leqslant \rho_{XX'} \qquad [2.27]$$

Por tanto, α puede considerarse una estimación del límite inferior del coeficiente de fiabilidad de un test.

Se han propuesto otros índices como estimaciones del límite inferior de la fiabilidad. Así, Lord y Novick (1968) recomiendan δ_3 de Guttman (1945), que según ellos da estimaciones al menos tan buenas como α y tiene las ventajas de ser siempre positivo, mientras que α puede ser negativo cuando hay correlaciones negativas entre los ítems, lo que lo invalida como estimador de la fiabilidad. El citado estimador viene dado por:

$$\delta_3 = 1 - \left(\sum_{j=1}^{n} \frac{\sigma_J^2}{\sigma_X^2} \right) + \frac{\sqrt{n/(n-1) \sum_{j \neq k}^{n} \sum \operatorname{cov}(J, K)}}{\sigma_X^2} \qquad [2.28]$$

donde los términos significan lo mismo que en la fórmula de α.

Por su parte, McDonald (1970, 1978, 1986) se muestra también crítico con α, negando que sea un buen indicador de la generalizabilidad, así como del límite inferior de la fiabilidad.

Cabe señalar que si bien el coeficiente alfa puede utilizarse tanto si los ítems son dicotómicos como si provienen de una escala ordinal con varias categorías, tipo Likert, en este último caso se han propuesto algunas variaciones que mejoran la estimación de la consistencia interna de la prueba respecto a la fórmula clásica aquí expuesta; pueden consultarse a tal efecto los trabajos de Elosua y Zumbo (2008) o Zumbo et al. (2007). Para el cálculo de α existen numerosos paquetes estadísticos, de libre acceso, como el software R (Elosua, 2009), o comerciales, como el SPSS, entre otros.

7.2. Casos particulares de α

Previamente a la presentación del coeficiente α por Cronbach en 1951, la psicometría clásica ya disponía de otras fórmulas para estimar la fiabilidad en términos de la consistencia interna del test. Dado que α constituye una solución más general al problema, se presentan aquí cuatro de esas fórmulas como casos particulares de α:

— Rulon (1939).
— Guttman (1945)/Flanagan (1937).
— Kuder y Richardson (1937):

- KR_{20}.
- KR_{21}.

Rulon

$$\rho_{XX'} = 1 - \frac{\sigma_d^2}{\sigma_X^2} \qquad [2.29]$$

donde:

σ_d^2: Varianza de las diferencias entre las puntuaciones de los sujetos en las dos mitades del test.
σ_X^2: Varianza de las puntuaciones globales de los sujetos en el test.

La fórmula de Rulon es una estimación de la fiabilidad del test a partir de las puntuaciones obtenidas en sus dos mitades, que se asumen paralelas, y, por tanto, las puntuaciones en ellas solo diferirán debido al error aleatorio. Nótese que [2.29] emerge directamente de la definición de coeficiente de fiabilidad dada en [2.2], $\rho_{XX'} = 1 - (\sigma_e^2/\sigma_X^2)$, al considerar, como se ha dicho, que la diferencia entre las dos mitades se debe únicamente al error, es decir, se define la varianza de los errores como la varianza de las diferencias.

Guttman-Flanagan

$$\rho_{XX'} = 2\left(1 - \frac{\sigma_p^2 + \sigma_i^2}{\sigma_X^2}\right) \qquad [2.30]$$

donde:

- σ_p^2: Varianza de las puntuaciones obtenidas por los sujetos en los ítems pares.
- σ_i^2: Varianza de las puntuaciones obtenidas en los ítems impares.
- σ_X^2: Varianza de las puntuaciones globales.

La fórmula de Guttman-Flanagan es equivalente a la de Rulon, expresando la varianza de las diferencias que aparecía en la fórmula de Rulon en función de las varianzas de la mitad par e impar del test. Puede tratar de demostrarlo el lector a modo de ejercicio.

Tanto [2.29] como [2.30] son casos particulares de α cuando $n = 2$, precisamente las dos mitades. En ese caso viene dada por:

$$\alpha = \frac{2}{2-1}\left(1 - \frac{\sigma_1^2 + \sigma_2^2}{\sigma_X^2}\right)$$

que es idéntica a [2.30], donde 2 y 1 se refieren a la mitad par e impar, respectivamente.

La estimación de la fiabilidad a partir de dos mitades resulta bastante problemática: ¿qué mitades tomar? Un test con n ítems tiene muchas posibles mitades, exactamente combinaciones de n elementos tomados de $n/2$ en $n/2$. Por ejemplo, un test con 10 ítems (solo 10) tiene 252 posibles mitades, o sea, por ese método se pueden hacer 126 estimaciones de su fiabilidad. Se demuestra (Cronbach, 1951) que α calculado a partir de todos los ítems de un test es el valor medio que se obtendría de calcularlo para todas las posibles mitades del test, es el valor esperado de las mitades: $\alpha = E(\alpha/2)$.

Kuder-Richardson

En su famoso artículo de 1937, Kuder y Richardson presentan, entre otras, sus no menos famosas fórmulas KR_{20} y KR_{21}, denominadas así por hacer precisamente los números 20 y 21 de las presentadas por los autores.

$$KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} p_j q_j}{\sigma_X^2}\right) \qquad [2.31]$$

KR_{20} es un caso particular de α cuando los ítems son dicotómicos, pues en ese caso, como es bien sabido, la varianza de una variable dicotómica viene dada por $\sigma_j^2 = p_j q_j$, siendo p_j la proporción de personas que aciertan el ítem j, y q_j, la proporción de los que lo fallan.

$$KR_{21} = \frac{n}{n-1}\left[1 - \frac{\bar{X} - (\bar{X}^2/n)}{\sigma_X^2}\right] \qquad [2.32]$$

KR_{21} es un caso particular de α cuando además de dicotómicos los ítems tienen la misma dificultad, en cuyo caso:

$$\sum_{j=1}^{n} p_j q_j = np_j q_j = npq = np(1-p) = np - npp =$$

$$= np - \frac{nppn}{n} = \bar{X} - \frac{\bar{X}^2}{n}$$

puesto que $np = \bar{X}$ cuando p es constante para todos los ítems.

Demuestre el lector, a modo de ejercicio, que si se utiliza KR_{21} con ítems cuya dificultad no es la misma para todos ellos, se obtiene un resultado menor que el que se obtendría utilizando KR_{20}.

Estas cuatro fórmulas que se acaban de citar tenían otrora una gran utilidad práctica, dada su sencillez para el cálculo; pero en la actualidad, en que todos los cálculos se hacen mediante ordenador, han caído en desuso y sencillamente se calcula α en su forma general.

Finalmente, veamos un ejemplo en el que se apliquen las fórmulas propuestas.

EJEMPLO

Un test que consta de seis ítems se aplicó a una muestra de cinco personas, obteniéndose los resultados de la tabla 2.1, en la que 1 significa acierto, y 0, fallo. Calcular α, Rulon, Guttman-Flanagan, KR_{20} y KR_{21}.

TABLA 2.1

Sujetos	Ítems									
	1	2	3	4	5	6	*X*	*P*	*I*	*P* − *I*
A	1	1	0	1	0	0	3	2	1	1
B	1	0	1	1	1	0	4	1	3	–2
C	0	1	1	0	0	0	2	1	1	0
D	1	1	1	1	1	1	6	3	3	0
E	1	0	0	0	0	0	1	0	1	–1

A la derecha de la tabla se han colocado las puntuaciones de las personas en los ítems pares (P), impares (I) y la diferencia ($P - I$).

Obtendremos previamente todos los datos necesarios para la aplicación de las fórmulas:

— Media total:

$$\bar{X} = \frac{3+4+2+6+1}{5} = 3{,}2$$

— Varianza:

$$S_X^2 = \frac{3^2+4^2+2^2+6^2+1^2}{5} - (3{,}2)^2 = 2{,}96$$

— Varianzas de los ítems:

$S_1^2 = (4/5)(1/5) = 0{,}16$ $\quad S_2^2 = (3/5)(2/5) = 0{,}24$

$S_3^2 = (3/5)(2/5) = 0{,}24$ $\quad S_4^2 = (2/5)(3/5) = 0{,}24$

$S_5^2 = (3/5)(2/5) = 0{,}24$ $\quad S_6^2 = (1/5)(4/5) = 0{,}16$

— Varianza de los pares:

$$S_p^2 = 1{,}04$$

— Varianza de los impares:

$$S_i^2 = 0{,}96$$

— Varianza de la diferencia:

$$S_d^2 = 1{,}039$$

$$\alpha = \frac{6}{6-1}\left(1 - \frac{0{,}16+0{,}24+0{,}24+0{,}24+0{,}24+0{,}16}{2{,}96}\right) =$$

$$= 0{,}681$$

$$\text{Rulon} = 1 - \frac{1{,}039}{2{,}96} = 0{,}648$$

$$\text{Guttman-Flanagan} = 2\left(1 - \frac{1{,}04+0{,}96}{2{,}96}\right) = 0{,}648$$

$$KR_{20} = \left(\frac{6}{6-1}\right)\left(1 - \frac{0{,}16+0{,}24+0{,}24+0{,}24+0{,}24+0{,}16}{2{,}96}\right) =$$

$$= 0{,}681$$

$$KR_{21} = \frac{6}{6-1}\left(1 - \frac{3{,}2+(3{,}2)^2/6}{2{,}96}\right) = 0{,}594$$

Nótese cómo efectivamente $KR_{21} < KR_{20}$ debido a que todos los ítems no tienen la misma dificultad. El uso de KR_{21} no estaría justificado en este caso.

7.3. Cálculo de α mediante análisis de varianza

Como Hoyt ya mostrara en 1941, el valor de α puede calcularse utilizando la técnica del análisis de varianza. No es difícil demostrar (véase, por ejemplo, Winer, 1971, pp. 283-296) que α viene dado por:

$$\alpha = \frac{nr_1}{1 + nr_1}$$

donde:

n: Número de ítems del test.
r_1: $[MC_{entre} - MC_{residual}]/nMC_{residual}$

siendo MC_{entre} la media cuadrática entre los sujetos y $MC_{residual}$ la media cuadrática residual, en un modelo de análisis de varianza de medidas repetidas, puesto que los ítems pueden ser considerados medidas repetidas de la misma variable, precisamente aquella que se trata de evaluar con el test.

Ejemplo

Utilizando el análisis de varianza, vamos a calcular el coeficiente α para los datos del ejemplo propuesto en el apartado anterior, comprobando que, efectivamente, de este modo se obtiene el mismo resultado que allí, es decir, $\alpha = 0{,}681$. Los datos del ejemplo eran los de la tabla 2.2.

TABLA 2.2

Sujetos	Ítems						
	1	2	3	4	5	6	Total sujetos
A	1	1	0	1	0	0	3
B	1	0	1	1	1	0	4
C	0	1	1	0	0	0	2
D	1	1	1	1	1	1	6
E	1	0	0	0	0	0	1
Total ítems	**4**	**3**	**3**	**3**	**2**	**1**	**16**

A partir de los datos de la tabla resulta muy sencillo obtener el valor de MC_{entre} y $MC_{residual}$. El lector poco familiarizado con el modelo de análisis de varianza de medidas repetidas puede consultar una exposición clara y detallada en Amón (1984) o Winer (1971), entre otros.

Pasos para el cálculo:

$$a) \quad \frac{T^2}{(n)(N)} = \frac{16^2}{(6)(5)} = 8{,}53333.$$

$$b) \quad \sum j^2 = 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 0^2 + 1^2 + 0^2 + \cdots + 0^2 + 0^2 + 0^2 = 16.$$

$$c) \quad \frac{\sum I^2}{N} = \frac{4^2 + 3^2 + 3^2 + 3^2 + 2^2 + 1^2}{5} = 9{,}6.$$

$$d) \quad \frac{\sum S^2}{n} = \frac{3^2 + 4^2 + 2^2 + 6^2 + 1^2}{6} = 11.$$

$$MC_{entre} = \frac{SC_{entre}}{N-1} = \frac{d-a}{N-1} = \frac{11 - 8{,}53333}{5-1} = 0{,}61666$$

$$MC_{residual} = \frac{SC_{residual}}{(n-1)(N-1)} = \frac{b-c-d+a}{(n-1)(N-1)} = \frac{16 - 9{,}6 - 11 + 8{,}53333}{(6-1)(5-1)} = 0{,}19666$$

$$r_1 = \frac{0{,}61666 - 0{,}19666}{6(0{,}19666)} = 0{,}35594$$

$$\alpha = \frac{6(0{,}35594)}{1 + 6(0{,}35594)} = 0{,}81$$

Que, efectivamente, es el resultado obtenido anteriormente utilizando la fórmula propuesta por Cronbach (1951).

7.4. Coeficiente beta (β)

Raju (1977) propuso el coeficiente β, generalización de α, que permite resolver un problema que se plantea cuando se dispone de una batería compuesta de varios subtest y se está interesado en obtener una estimación del coeficiente α de la batería basándose en los datos de los subtest considerados como componentes, al modo de los ítems de un test, es decir, aquí los subtest serían los ítems de la batería. Pues bien, en esas circunstancias, si los subtest tienen distinto número de ítems, que suele ser lo más frecuente, y se calcula el coeficiente α global de la batería, la estimación resulta ser una infraestimación, y ese es precisamente el problema que viene a resolver Raju. En suma, el coeficiente β evita el problema de la infraestimación de α en las baterías cuando los subtest que las componen tienen distinto número de ítems, siendo igual a α cuando el número de ítems de los subtest es el mismo.

Ha de tenerse claro que todo esto tiene sentido si se desconocen los datos directos de las respuestas de las personas a cada ítem, pues, conocidas estas, es más recomendable calcular α directamente, considerando la batería un test formado por los ítems de todos los subtest. No obstante, puede haber situaciones, de hecho hay bastantes, en las que solo se dispone de los datos referidos a los subtest, en cuyo caso β resulta apropiado.

El coeficiente β viene dado por la fórmula:

$$\beta = \frac{\sigma_X^2 - \sum_{j=1}^{k} \sigma_j^2}{\sigma_X^2 \left[1 - \sum_{j=1}^{k} \left(\frac{n_j}{n}\right)^2\right]} \qquad [2.33]$$

donde:

K: Número de subtest de la batería.
σ_X^2: Varianza de las puntuaciones globales de la batería.
σ_j^2: Varianza de cada subtest.
n_j: Número de ítems de cada subtest.
n: Número de ítems total de la batería.

Ejemplo

Una batería de procesos básicos está formada por tres subtest: memoria a corto plazo (MCP), memoria a largo plazo (MLP) y tiempo de reacción (TR), con 10, 25 y 40 ítems, respectivamente. Aplicada la batería a una muestra, se encontró una varianza de las puntuaciones globales igual a 32, siendo las varianzas de los subtest 6, 8 y 10, respectivamente.

Calcular el coeficiente α y el coeficiente β de la batería en función de los subtest.

$$\alpha = \frac{3}{3-1}\left(1 - \frac{6+8+10}{32}\right) = 0{,}375$$

$$\beta = \frac{32 - (6+8+10)}{32[1 - \{(10/75)^2 + (25/75)^2 + (40/75)^2\}]} = 0{,}426$$

El coeficiente α estimado según la corrección de Raju pasa de 0,375 a 0,426.

Nótese que, efectivamente, si el número de ítems de los subtest es el mismo, entonces $\alpha = \beta$, pues:

$$N = Kn_j$$

luego

$$1 - \sum_{j=1}^{k}\left(\frac{n_j}{n}\right)^2 = 1 - \sum_{j=1}^{k}\left(\frac{n_j}{Kn_j}\right)^2 = 1 - \sum_{j=1}^{k}\frac{1}{K^2} =$$

$$= 1 - \frac{K}{K^2} = 1 - \frac{1}{K} = \frac{K-1}{K}$$

luego

$$\beta = \frac{\sigma_X^2 - \sum_{j=1}^{k}\sigma_j^2}{\sigma_X^2(K-1)/K} = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{k}\sigma_j^2}{\sigma_X^2}\right) = \alpha$$

7.5. Coeficientes basados en el análisis factorial de los ítems

A partir de los datos proporcionados por el análisis factorial de los ítems de un test se pueden obtener indicadores de la consistencia interna muy semejantes al coeficiente α. Ni que decir tiene que el propio resultado del análisis factorial ya constituye un excelente indicador de la consistencia interna de los ítems, analizando la matriz de correlaciones, el número de factores obtenidos y la varianza explicada por cada uno de ellos. No obstante, tiene interés la obtención de algún índice único que sintetice de una forma razonable toda esta información.

Dos de los índices más usados son la theta (θ) de Carmines (Carmines y Zeller, 1979) y la omega (Ω) de Heise y Bohrnstedt (1970).

Coeficiente theta (θ)

La fórmula del coeficiente θ viene dada por:

$$\theta = \frac{n}{n-1}\left(1 - \frac{1}{\lambda_1}\right)$$

donde:

- n: Número de ítems.
- λ_1: Valor propio *(eigenvalue)* mayor, es decir, la varianza explicada por el primer factor antes de la rotación.

La interpretación de θ es clara: cuanta más varianza explica el primer factor, mayor será θ, lo que indicará que los ítems están más intercorrelacionados y tienden a articularse en torno a una sola dimensión. θ puede, por tanto, ser considerado un indicador del grado de unidimensionalidad de los ítems. Como señalan Carmines y Zeller (1979), θ constituye una buena estimación del límite superior de α, es decir, $\alpha \leqslant \theta$.

Ejemplo

En la tabla 2.3 aparece la varianza explicada *(Eigenvalues)* por los 15 factores obtenidos al someter a un análisis factorial 15 variables (Muñiz, García y Virgós, 1991). Calcular θ.

TABLA 2.3

Factor	Varianza explicada *(Eigenvalues)*
1	9,9564
2	1,3327
3	0,9839
4	0,6498
5	0,4976
6	0,3751
7	0,3230
8	0,2567
9	0,1931
10	0,1557
11	0,0882
12	0,0669
13	0,0544
14	0,0340
15	0,0325

Sustituyendo en la fórmula:

$$\theta = \frac{15}{15-1}\left(1 - \frac{1}{9{,}9564}\right) = 0{,}9638$$

Para los mismos datos, el valor de α resultó ser de 0,9613, comprobándose que, efectivamente, su valor es inferior que el de θ.

Coeficiente omega (Ω)

Su fórmula viene dada por:

$$\Omega = 1 - \frac{\sum \sigma_i^2 - \sum \sigma_i^2 h_i^2}{\sum_{j \neq 1} \sum \sigma_{ij}}$$

donde:

- $\Sigma\sigma_i^2$: Suma de las varianzas de los ítems.
- h_i^2: Comunalidad estimada del ítem i.
- $\Sigma\Sigma\sigma_{ij}$: Suma de las covarianzas entre los ítems.

La fórmula anterior puede expresarse de un modo más sencillo en función de las correlaciones entre los ítems:

$$\Omega = 1 - \frac{n - \sum h_i^2}{n + 2\sum r_{ij}}$$

donde:

Σr_{ij}: Suma de las correlaciones entre los ítems, y el resto de los términos son los de la fórmula anterior.

Para los datos del ejemplo anterior, Ω viene dado por:

$$\Omega = 1 - \frac{15 - 11{,}94898}{15 + 2(66{,}623)} = 0{,}9791$$

Nótese que, como señalan Armor (1974) y Carmines y Zeller (1979), el valor de Ω para unos mismos datos es superior al de α y θ. En el caso de que los ítems fuesen paralelos, los tres coeficientes serían iguales entre sí, y además serían iguales al coeficiente de fiabilidad del test. De lo contrario, su cuantía viene dada en el siguiente orden: $\alpha < \theta < \Omega$. Por tanto, de los tres estimadores de la consistencia interna que acabamos de exponer, α es el que proporciona valores ligeramente más bajos.

7.6. Inferencias sobre α

Hasta ahora se ha visto únicamente cómo se calcula el coeficiente α en una muestra, pero, una vez hallado su valor, seguramente se le plantearán al investigador o profesional varias preguntas, por ejemplo, si su coeficiente es estadísticamente significativo, cuál es el valor de α en la población, o si la diferencia entre su coeficiente α y el hallado por otro investigador es significativa, etc. De este tipo de cuestiones, y otras parecidas, concernientes a los aspectos inferenciales acerca de α es de lo que trata el presente apartado.

La teoría del error muestral para el coeficiente α ha sido desarrollada por diversos autores (Feldt, 1965, 1969, 1980; Hakstian y Whalen, 1976; Kristof, 1963, entre otros). Aquí seguiremos básicamente la exposición sistemática y clara realizada por Feldt, Woodruff y Salih (1987). Se considerarán tres casos:

a) Inferencias acerca de un solo coeficiente.
b) Comparación de coeficientes obtenidos en muestras independientes.
c) Comparación de coeficientes obtenidos en la misma muestra.

Un solo coeficiente

Una vez calculado el valor de α en una muestra ($\hat{\alpha}$), cabe preguntarse si a un determinado nivel de confianza el valor obtenido es compatible con la hipótesis de que α tome determinado valor en la población, siendo especialmente habitual preguntarse si es compatible con la hipótesis de que su valor en la población sea cero, es decir, preguntarse si resulta estadísticamente significativo. Kristof (1963) y Feldt (1965) han propuesto a tal efecto (véase apéndice) el estadístico de contraste:

$$F = \frac{1 - \alpha}{1 - \hat{\alpha}} \qquad [2.34]$$

que se distribuye según F con $(N - 1)$ y $(n - 1)(N - 1)$ grados de libertad, y donde:

N: Número de casos de la muestra.
n: Número de ítems del test.
α: Valor de alfa en la población.
$\hat{\alpha}$: Valor de alfa en la muestra.

Ejemplo

Un test de inteligencia espacial que constaba de 50 ítems se aplicó a una muestra de 40 personas, obteniéndose un coeficiente $\hat{\alpha} = 0{,}75$. Utilizando el nivel de confianza del 95%:

1. ¿Puede afirmarse que el coeficiente encontrado es estadísticamente significativo?
2. ¿Entre qué valores se estima que se encontrará el coeficiente α de la población?

1. Hipótesis nula:

$$H_0: \ \alpha = 0$$

Hipótesis alternativa:

$$H_1: \ \alpha \neq 0$$

Calculamos el valor del estadístico de contraste:

$$F = \frac{1-\alpha}{1-\hat{\alpha}} = \frac{1-0}{1-0{,}75} = 4$$

Los valores críticos de F en las tablas (bilateral) vienen dados por:

$$F_{0{,}975(39{,}1911)} = 1{,}48$$
$$F_{0{,}025(39{,}1911)} = 0{,}61$$

Dado que el valor del estadístico de contraste ($F = 4$) no está comprendido entre los valores críticos 0,61 y 1,48, se rechaza la hipótesis nula al NC del 95%, y se puede afirmar, por tanto, que el coeficiente α encontrado es estadísticamente significativo, es decir, es incompatible con la hipótesis de que el valor de α en la población sea 0.

2. Para resolver la segunda pregunta basta con sustituir los valores críticos de F ya calculados en el estadístico de contraste y despejar α:

$$\frac{1-\alpha}{1-0{,}75} \leqslant 1{,}48 \quad \text{de donde} \quad \alpha \geqslant 0{,}63$$

$$\frac{1-\alpha}{1-0{,}75} \geqslant 0{,}61 \quad \text{de donde} \quad \alpha \leqslant 0{,}8475$$

Por tanto:

$$0{,}63 \leqslant \alpha \leqslant 0{,}8475$$

En suma, al NC del 95%, α estará comprendido entre 0,63 y 0,8475. Nótese que la hipótesis de significación estadística planteada en la primera pregunta puede resolverse a la luz del intervalo confidencial ahora calculado, dado que si se estima que el α de la población estará entre 0,63 y 0,8475, ello quiere decir que el valor 0 de la hipótesis nula no cae dentro de dicho intervalo; luego, como efectivamente se hizo allí, se rechaza dicha hipótesis y se afirma la significación estadística. Lo cual puede resumirse diciendo que el problema de la significación estadística es un caso particular de la comprobación de hipótesis; no obstante se mantendrá la distinción terminológica por ser tan habitual en la literatura psicológica.

Es evidente que, una vez establecidos los límites entre los que se considera que se encuentra el valor de α en la población, quedan resueltas automáticamente todas las hipótesis acerca de su cuantía.

Muestras independientes

a) Dos coeficientes

Feldt (1969) propuso un estadístico de contraste que permite comparar dos coeficientes α obtenidos en muestras independientes:

$$w = \frac{1-\hat{\alpha}_1}{1-\hat{\alpha}_2} \qquad [2.35]$$

donde w se distribuye según F con $(N_1 - 1)$ y $(N_2 - 1)$ grados de libertad (véase apéndice), y $\hat{\alpha}_1$ y $\hat{\alpha}_2$ son los valores de α en cada muestra.

Ejemplo

El coeficiente α de una escala de dogmatismo en una muestra de 55 mujeres fue 0,60 y en una muestra de 126 hombres resultó ser 0,67. Al NC del 95%, ¿puede afirmarse que existen diferencias estadísticamente significativas entre ambos coeficientes?

$$H_0: \alpha_1 = \alpha_2$$
$$H_1: \alpha_1 \neq \alpha_2$$
$$w = \frac{1-0{,}60}{1-0{,}67} = 1{,}21$$

En las tablas de F:

$$F_{0{,}975(54.125)} = 1{,}53$$
$$F_{0{,}025(54.125)} = 0{,}63$$

Como $w = 1{,}21$ se admite la hipótesis nula, la diferencia no resulta estadísticamente significativa al NC del 95%.

b) K coeficientes

Woodruff y Feldt (1986) derivaron el estadístico de contraste UX_1 que permite la comparación si-

multánea de K coeficientes α obtenidos en muestras independientes. Por su parte, Hakstian y Whalen (1976) propusieron un método ligeramente distinto a este que conduce a resultados similares.

$$UX_1 = \frac{\sum_{i=1}^{k}[(1-\hat{\alpha}_i)^{-1/3} - \bar{u}]^2}{\bar{S}^2} \qquad [2.36]$$

donde:

K: Número de muestras independientes.
UX_1: Se distribuye según χ^2 (muy aproximadamente) con $K-1$ grados de libertad.
$\hat{\alpha}_i$: Valor de α en cada muestra i.
$\bar{u}$: Viene dado por:

$$\sum_{i=1}^{k}\frac{(1-\hat{\alpha}_i)^{-1/3}}{K}$$

$\bar{S}$: Viene dado por:

$$\bar{S}^2: \sum_{i=1}^{k}\frac{S_i^2}{K}$$

donde

$$S_i^2 = \frac{2}{9(\tilde{N}_i - 1)(1-\alpha_i)^{2/3}}$$

y

$$\tilde{N}_i = \frac{N_i(n_i - 1)}{n_i + 1}$$

siendo

N_i: Número de personas de cada muestra.
n_i: Número de ítems de cada test.

Ejemplo

Para evaluar la depresión se utilizaron cuatro versiones de un cuestionario (D_1, D_2, D_3, D_4) aplicada cada una de ellas a una muestra independiente de personas. Cada escala constaba de 25 ítems, y el número de personas de las muestras fue, respectivamente, de 100, 70, 80 y 104. Se obtuvieron para α los siguientes valores:

$$\hat{\alpha}_1 = 0{,}7,\ \hat{\alpha}_2 = 0{,}5,\ \hat{\alpha}_3 = 0{,}6,\ \hat{\alpha}_4 = 0{,}8$$

Al nivel de confianza del 95 %, ¿puede afirmarse que las diferencias entre los valores de α obtenidos son estadísticamente significativas?

$$H_0:\ \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$$

$$\bar{u} = \frac{(1-0{,}7)^{-1/3}}{4} + \frac{(1-0{,}5)^{-1/3}}{4} + \frac{(1-0{,}6)^{-1/3}}{4} + \frac{(1-0{,}8)^{-1/3}}{4} = 1{,}455$$

$$\tilde{N}_1 = \frac{100(25-1)}{25+1} = 92{,}31$$

$$\tilde{N}_2 = \frac{70(25-1)}{25+1} = 64{,}61$$

$$\tilde{N}_3 = \frac{80(25-1)}{25+1} = 73{,}85$$

$$\tilde{N}_4 = \frac{104(25-1)}{25+1} = 96{,}00$$

$$S_1^2 = \frac{2}{9(92{,}31-1)(1-0{,}7)^{2/3}} = 0{,}0054$$

$$S_2^2 = \frac{2}{9(64{,}61-1)(1-0{,}5)^{2/3}} = 0{,}0055$$

$$S_3^2 = \frac{2}{9(73{,}85-1)(1-0{,}6)^{2/3}} = 0{,}0056$$

$$S_4^2 = \frac{2}{9(96{,}00-1)(1-0{,}8)^{2/3}} = 0{,}0068$$

$$\bar{S}^2 = \frac{0{,}0054 + 0{,}0055 + 0{,}0056 + 0{,}0068}{4} = 0{,}0058$$

$$UX_1 = \frac{[(1-0{,}7)^{-1/3} - 1{,}455]^2}{0{,}0058} + \frac{[(1-0{,}5)^{-1/3} - 1{,}455]^2}{0{,}0058} + \frac{[(1-0{,}6)^{-1/3} - 1{,}455]^2}{0{,}0058} + \frac{[(1-0{,}8)^{-1/3} - 1{,}455]^2}{0{,}0058} = 19{,}679$$

Al NC del 95% los valores críticos de χ^2 (bilateral) con 3 grados de libertad ($K - 1 = 4 - 1 = 3$) son 0,22 y 9,35. Dado que nuestro valor de 19.679 no está comprendido en dicho intervalo, rechazamos la hipótesis nula de igualdad entre los coeficientes de fiabilidad α. Nótese que la hipótesis que se rechaza es que todos sean iguales; sin embargo, podría ocurrir, por ejemplo, que dos de ellos sí lo fuesen. Para hacer comparaciones por pares puede utilizarse el estadístico de contraste expuesto en [2.35] o este mismo UX_1 con $K = 2$. Veamos en el caso de nuestro ejemplo, a modo de ilustración, si se puede afirmar que $\alpha_4 = 0{,}8$ resulta estadísticamente superior a $\alpha_2 = 0{,}5$.

$$H_0: \alpha_4 = \alpha_2$$
$$H_1: \alpha_4 > \alpha_2$$

$$\bar{u} = \frac{(1 - 0{,}5)^{-1/3}}{2} + \frac{(1 - 0{,}8)^{-1/3}}{2} = 1{,}485$$

$$S_2^2 = 0{,}0055$$

$$S_4^2 = 0{,}0068$$

$$\bar{S}^2 = \frac{0{,}0055 + 0{,}0068}{2} = 0{,}00615$$

$$UX_1 = \frac{[(1 - 0{,}5)^{-1/3} - 1{,}485]^2}{0{,}00615} + \frac{[(1 - 0{,}8)^{-1/3} - 1{,}485]^2}{0{,}00615} =$$

$$= 16{,}47$$

Al nivel de confianza del 95%, el valor crítico de χ^2 (unilateral) con 1 grado de libertad ($K - 1 = 2 - 1 = 1$) viene dado por 3,84; por tanto, dado que $16{,}47 > 3{,}84$, se rechaza la hipótesis nula de igualdad, afirmándose al NC del 95% que α_4 es estadísticamente superior a α_2.

Al mismo resultado se llegaría utilizando el estadístico de contraste expuesto en [2.35]:

$$w = \frac{1 - 0{,}5}{1 - 0{,}8} = 2{,}5$$

El valor crítico correspondiente a $F_{0{,}95(69,103)}$ (unilateral) en las tablas es 1,43 menor que 2,5; luego, como antes, se rechaza la hipótesis nula de igualdad.

Muestras dependientes

a) *Dos coeficientes*

Los investigadores y profesionales disponen a menudo de los coeficientes α de dos test que han sido calculados en muestras dependientes; tal es el caso cuando ambos coeficientes han sido calculados en la misma muestra. Feldt (1980) propuso un estadístico de contraste que permite comparar dos coeficientes α obtenidos en la misma muestra:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N - 2}}{\sqrt{4(1 - \hat{\alpha}_1)(1 - \hat{\alpha}_2)(1 - \hat{\rho}_{12}^2)}} \qquad [2.37]$$

donde:

- t: Se distribuye con $N - 2$ grados de libertad.
- N: Número de sujetos de la muestra.
- $\hat{\alpha}_1$ y $\hat{\alpha}_2$: Valores de los coeficientes α de los test.
- $\hat{\rho}_{12}$: Correlación entre las puntuaciones de las personas en ambos test.

Ejemplo

Dos test de independencia de campo se aplicaron a una muestra de 227 personas, obteniéndose una correlación entre las puntuaciones de las personas en ambos de 0,6. Los coeficientes α para cada uno de los test fueron, respectivamente, 0,70 y 0,85. Al nivel de confianza del 95%, ¿puede afirmarse que la diferencia entre ambos coeficientes α es estadísticamente significativa?

$$H_0: \alpha_1 = \alpha_2$$
$$H_1: \alpha_1 \neq \alpha_2$$

$$t = \frac{(0{,}85 - 0{,}70)\sqrt{227 - 2}}{\sqrt{4(1 - 0{,}85)(1 - 0{,}70)(1 - 0{,}6^2)}} = 6{,}63$$

Los valores críticos de t en las tablas (bilateral) con 225 grados de libertad vienen dados por –1,972 y +1,972. Dado que el valor obtenido (6,63) cae fuera de dicho intervalo, se afirma que al NC del

95% la diferencia entre los coeficientes es estadísticamente significativa; se rechaza, por tanto, la hipótesis nula.

b) *K coeficientes*

Para comparar simultáneamente más de dos coeficientes α obtenidos en la misma muestra, Woodruff y Feldt (1986) propusieron el siguiente estadístico de contraste:

$$UX_2 = \frac{\sum_{i=1}^{k}[(1-\hat{\alpha}_i)^{-1/3} - \overline{u}]^2}{\overline{S}^2 - \overline{S}_{jk}} \qquad [2.38]$$

donde:

K: Número de test.
N: Número de personas de la muestra.
UX_2: Se distribuye (aproximadamente) según χ^2 con $K-1$ grados de libertad.
$\hat{\alpha}_i$: Valor empírico de los coeficientes en la muestra.

$\overline{u}$: $\sum_{i=1}^{k} \frac{1}{K(1-\hat{\alpha}_i)^{1/3}}$

$\overline{S}^2$: $\sum_{i=1}^{k} \frac{S_i^2}{k}$

con

S_i^2: $\frac{2}{9(\tilde{N}-1)(1-\hat{\alpha}_i)^{2/3}}$

$\tilde{N}$: $\frac{N(\tilde{n}-1)}{\tilde{n}+1}$

$\tilde{n}$: $\frac{K}{\sum_{i=1}^{k}\frac{1}{n_i}}$ (media armónica de la longitudes de los test)

S_{jk}: $\frac{2\hat{\rho}_{jk}^2}{9(\tilde{N}-1)(1-\hat{\alpha}_j)^{1/3}(1-\hat{\alpha}_k)^{1/3}}$

$\overline{S}_{jk}$: $\frac{\sum S_{jk}}{K(K-1)/2}$

Ejemplo

Se aplicaron cuatro test a una muestra de 200 personas, obteniéndose las siguientes correlaciones entre las puntuaciones:

$$r_{12} = 0{,}40, \quad r_{13} = 0{,}50, \quad r_{14} = 0{,}60$$
$$r_{23} = 0{,}45, \quad r_{24} = 0{,}55, \quad r_{34} = 0{,}30$$

Los coeficientes α de cada uno de ellos fueron, respectivamente: 0,70, 0,75, 0,80 y 0,85. Los test constaban de 20, 25, 30 y 35 ítems, respectivamente. Al nivel de confianza del 95%, ¿puede afirmarse que existen diferencias estadísticamente significativas entre los coeficientes α de los cuatro test?

$$H_0: \ \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$$

$$K = 4$$

$$\overline{u} = \frac{1}{4}(1-0{,}70)^{-1/3} + \frac{1}{4}(1-0{,}75)^{-1/3} + \frac{1}{4}(1-0{,}80)^{-1/3} + \frac{1}{4}(1-0{,}85)^{-1/3} = 1{,}67$$

$$\tilde{n} = \frac{4}{(1/20)+(1/25)+(1/30)+(1/35)} = 26{,}33$$

$$\tilde{N} = \frac{200(26{,}33-1)}{(26{,}33+1)} = 185{,}36$$

$$S_1^2 = \frac{2}{9(185{,}36-1)(1-0{,}70)^{2/3}} = 0{,}00269$$

$$S_2^2 = \frac{2}{9(185{,}36-1)(1-0{,}75)^{2/3}} = 0{,}00304$$

$$S_3^2 = \frac{2}{9(185{,}36-1)(1-0{,}80)^{2/3}} = 0{,}00352$$

$$S_4^2 = \frac{2}{9(185{,}36-1)(1-0{,}85)^{2/3}} = 0{,}00427$$

$$\overline{S}^2 = \frac{0{,}00269 + 0{,}00304 + 0{,}00352 + 0{,}00427}{4} = 0{,}00338$$

$$S_{12} = \frac{2(0{,}40)}{9(185{,}36-1)(1-0{,}70)^{1/3}(1-0{,}75)^{1/3}} = 0{,}00114$$

$$S_{12} = \frac{2(0,40)}{9(185,36-1)(1-0,70)^{1/3}(1-0,75)^{1/3}} = 0,00114$$

$$S_{13} = \frac{2(0,50)}{9(185,36-1)(1-0,70)^{1/3}(1-0,80)^{1/3}} = 0,00154$$

$$S_{14} = \frac{2(0,60)}{9(185,36-1)(1-0,70)^{1/3}(1-0,85)^{1/3}} = 0,00203$$

$$S_{23} = \frac{2(0,45)}{9(185,36-1)(1-0,75)^{1/3}(1-0,80)^{1/3}} = 0,00147$$

$$S_{24} = \frac{2(0,55)}{9(185,36-1)(1-0,75)^{1/3}(1-0,85)^{1/3}} = 0,00198$$

$$S_{34} = \frac{2(0,30)}{9(185,36-1)(1-0,80)^{1/3}(1-0,85)^{1/3}} = 0,00110$$

$$\overline{S}_{jk} = \frac{0,00114 + 0,00154 + 0,00203 + 0,00147}{4(4-1)/2} + \frac{0,00198 + 0,00110}{4(4-1)/2} = 0,00154$$

$$UX_2 = \frac{[(1-0,70)^{-1/3} - 1,67]^2}{0,00338 - 0,00154} + \frac{[(1-0,75)^{-1/3} - 1,67]^2}{0,00338 - 0,00154} + \frac{[(1-0,80)^{-1/3} - 1,67]^2}{0,00338 - 0,00154} + \frac{[(1-0,85)^{-1/3} - 1,67]^2}{0,00338 - 0,00154} = 45,89$$

Los valores críticos de χ^2 con 3 grados de libertad ($K - 1 = 4 - 1 = 3$) vienen dados por 0,22 y 9,35. Dado que el valor hallado, 45,89, no cae dentro de dicho intervalo, se rechaza la H_0 o, lo que es lo mismo, se afirma al NC del 95% que existen diferencias estadísticamente significativas entre los coeficientes α. Nótese que la hipótesis alternativa no es que todos los coeficientes α sean desiguales, sino que no todos son iguales.

8. TEORÍA DE LA GENERALIZABILIDAD

8.1. Fuentes de error

Como se acaba de ver en las páginas precedentes, la teoría clásica define el error aleatorio de las mediciones como la diferencia entre la puntuación empírica de la persona (X) y su puntuación verdadera (V):

$$e = X - V$$

Es decir, el que una persona no obtenga la misma puntuación empírica en dos formas paralelas de un test, o en dos aplicaciones sucesivas del mismo test, se debe a que han intervenido ciertos factores distorsionadores que generan error aleatorio. Una medida será tanto más fiable cuanto menos error aleatorio contenga, cantidad que se estima mediante el coeficiente de fiabilidad. Ahora bien, ¿no sería posible penetrar dentro de ese error y averiguar a qué se debe exactamente?, ¿diseccionar los distintos factores que lo componen?; en suma, ¿no sería posible descubrir las fuentes de las que mana el error? Algo de error aleatorio incontrolable siempre habrá; todas las ciencias lo asumen en sus mediciones, pero sería deseable conocer de dónde proviene el grueso del error para así mejor evitarlo. La teoría clásica ha peleado con este asunto desde siempre, y su estrategia, que ha probado ser eficaz, ha sido la de mantener fijas todas las condiciones intervinientes en el proceso de medición, para así atribuir el error existente a variaciones espurias, que se espera sean mínimas. El planteamiento es correcto cuando se puede llevar a cabo; lo que ocurre es que en muchas situaciones de medición en las ciencias sociales esta fijación no se puede realizar, y es entonces cuando interesa saber cuánto error se debe a cada uno de los factores intervinientes. Por ejemplo, si se utiliza el método test-retest para calcular el coeficiente de fiabilidad, se asume que todo permanece igual: la competencia de las personas en la variable medida, el test, el aplicador, el lugar, etc. Por tanto, es razonable asumir que las variaciones entre una y otra aplicación se deban a ligeras variaciones de causa ignota y aleatoria. Sin embargo, hay numerosas situaciones en las que no todos los factores intervinientes son constantes, y en esos casos tiene gran interés averiguar qué parte del error aleatorio se debe a unos y a otros. Por ejemplo, si varios profesores evalúan a un grupo de niños con varios métodos de evaluación, existen dos fuentes potenciales de error:

a) Los profesores, ya que no todos ellos actuarán exactamente igual, es decir, la fiabilidad interprofesor no será perfecta.

b) Los métodos de evaluación, pues no será lo mismo si se trata de una prueba objetiva, o de un ensayo, que sea oral o escrita, etc.

Siempre quedará un cierto error aleatorio que no sabremos a qué atribuir, pero en este caso tenemos identificadas al menos dos fuentes potenciales de error. Claro, se podrá decir, lo que ocurre es que las cosas no se deben hacer así, habría que unificar profesores y métodos para eliminar su incidencia. Desde luego esa es una opción, incluso se podría decir que deseable; lo que ocurre es que no siempre es posible en las ciencias sociales, por lo que hay que disponer de tecnología psicométrica para atacar el problema cuando se presente. En el campo de la psicología de las aptitudes y de la personalidad es más frecuente la tendencia a unificar las condiciones de medida, pero en planteamientos de carácter más observacional y en muchas situaciones educativas se presentan numerosos casos en los que esto no es posible.

Analizar y descifrar racionalmente las fuentes de error no resulta especialmente complicado cuando se conoce a fondo la situación concreta de medición; lo que ya es algo más difícil es estimar la cuantía de los errores atribuible a esas fuentes. Empecemos por lo más sencillo: autores clásicos como Cronbach (1947), Thorndike (1951) o Stanley (1971) llevaron a cabo análisis racionales exhaustivos de las distintas fuentes del error.

El primer y más obvio manantial de errores es la *propia persona* que realiza el test. Su situación específica en ese momento (salud, humor, fatiga, motivación, etc.), su suerte, entrenamiento previo, acontecimientos personales recientes, etc., influyen para que su puntuación empírica en una prueba fluctúe de una a otra vez. Raramente podremos estimar la incidencia de este tipo de errores, todo lo más que se puede hacer es realizar la prueba en unas condiciones óptimas que los minimicen.

Las características del *instrumento de medida* utilizado, tales como ítems, formato, modo de respuesta, etc., también pueden incidir en las puntuaciones. Cuando el instrumento de medida, o alguno de sus ítems, interacciona con las personas, se habla de sesgo. Si no resulta idéntico para todas las personas a las que se aplica, está sesgado contra cierto tipo de ellas. Cuando se utiliza más de un instrumento hay que tratar de estimar los posibles errores introducidos por ello, y, en cualquier caso, se debe estudiar siempre la posible existencia de sesgo.

Los *aplicadores* pueden introducir sin pretenderlo distorsiones en los resultados, según lleven a cabo las instrucciones, el tipo de relación que establezcan con las personas evaluadas *(rapport),* su apariencia externa, características personales, etc. Está claro que si no están perfectamente entrenados, los aplicadores pueden no actuar homogéneamente. Por tanto, cuando se utilicen varios aplicadores, estos pueden introducir un cierto error de medida que conviene estimar.

Las *condiciones de aplicación,* tales como lugar, hora, día, características físicas (ruido, luz, visibilidad, etc.), pueden tener una influencia notable. Un ejemplo muy típico consiste en aplicar los test a los niños a última hora de la mañana o de la tarde, por conveniencias de horarios de clases, cuando los niños a esas horas lo único que desean es irse. No digamos nada si para aplicar los test se suprime alguna actividad de su natural agrado, tal como recreo, deportes, etc.

Acontecimientos nacionales e internacionales importantes pueden condicionar los resultados, especialmente si de algún modo interaccionan con el objetivo de la medida.

Finalmente, a los factores anteriores habría que añadir las posibles *interacciones* entre ellos. Por ejemplo, podría ocurrir que la pericia de los aplicadores no fuese igual con todos los test, haciéndolo mejor con un tipo de test que con otros; en ese caso hablaríamos de una interacción aplicador × instrumento.

A la vista de tantos factores potencialmente distorsionantes, resulta asombroso que los test tengan un coeficiente de fiabilidad razonable. Pero tampoco conviene alarmarse: si la variable medida tiene entidad, el instrumento utilizado para medirla está bien construido y se aplica adecuadamente, los errores son mínimos.

Si se planifica la medición de tal modo que se tengan en cuenta algunos de los factores citados, u otros cualesquiera, se podrá estimar su contribución al tamaño de los errores. Por ejemplo, si sospechamos que los entrevistadores influyen en el diagnóstico y también lo hace el tipo de entrevista, podremos diseñar una recogida de datos en la que se contemplen estos dos factores, para poder estimar posteriormente sus efectos.

Lo dicho hasta ahora acerca de la descomposición del error en sus distintos componentes solo implica que el investigador decida qué factores son los relevantes para su medición, para así tenerlos en cuenta. El problema será cómo estimar la cuantía del error debido a cada uno de los diferentes factores intervinientes. Para llevar a cabo esta tarea se han sugerido y formulado diversos modelos (Lord y Novick, 1968; Novick, 1966), pero el acercamiento más sistemático y ambicioso es el de la teoría de la generalizabilidad (TG). La TG hunde sus raíces en trabajos pioneros como los de Hoyt (1941), Burt (1955) o Lindquist (1953), aunque sus ideas y conceptos hallábanse latentes y dispersos por la literatura psicométrica clásica. Serán Cronbach y sus colaboradores (Cronbach et al., 1963; Gleser et al., 1965) quienes lleven a cabo una formulación sistemática, especialmente en su enciclopédico trabajo de 1972 (Cronbach, Gleser, Nanda y Rajaratnam, 1972). Exposiciones más asequibles pueden consultarse en Brennan (1983) y Shavelson y Webb (1991). Para una buena introducción, véanse Crocker y Algina (1986) o Shavelson, Webb y Rowley (1989); consejos útiles para su uso en la práctica pueden verse en Briesch et al. (2014), y un tratamiento completo, en Brennan (2001).

8.2. Conceptos básicos

El objetivo central de la TG es determinar las distintas fuentes de error que afectan a las mediciones y estimar su cuantía. Veamos cuáles son los conceptos clave de los que parte para llevar a cabo esta tarea. Lo haremos mediante un ejemplo bastante artificial, pero útil para iniciarse en los conceptos básicos.

Ejemplo

Supongamos una *población* de 10 personas que van a ser evaluadas para acceder a un puesto de trabajo por una *población* de cinco evaluadores. Cumplida su labor, las puntuaciones asignadas por los evaluadores a las personas en una escala de cero a diez puntos aparecen en la tabla 2.4 (la artificiosidad del ejemplo proviene de que lo normal es que las poblaciones de personas y evaluadores sean mucho más numerosas, suele asumirse que infinitas; por tanto, en la realidad se trabaja con muestras, no con poblaciones).

TABLA 2.4

		Población de evaluadores					
		E_1	E_2	E_3	E_4	E_5	μ_p
Población de personas	*a*	7	8	7	9	8	7,8
	b	4	4	2	3	4	3,4
	c	0	0	1	2	1	0,8
	d	6	6	5	4	5	5,2
	e	3	3	2	2	3	2,6
	f	2	2	2	2	2	2,0
	g	7	8	6	6	7	6,8
	h	4	5	5	4	4	4,4
	i	2	4	3	4	3	3,2
	j	9	9	8	7	8	8,2
	μ_i	4,4	4,9	4,1	4,3	4,5	**4,44**

X_{pi}: Es la *puntuación empírica* de una persona (p) en un instrumento de medida (i). Por ejemplo, en la tabla 2.4 la persona *g* tiene una puntuación $X_{pi} = 8$ con el evaluador 2, etc.

μ: Se denomina *gran media* y es la media de todas las personas de la población en todos los instrumentos de medida de la población de instrumentos que se contemple. En nuestro caso su valor es 4,44, que proviene de

hacer la media de las puntuaciones de todas las personas para todos los evaluadores.

μ_p: *Puntuación universo* de una persona *p.* Es la media de las puntuaciones de una persona determinada en todos los instrumentos de la población considerada. Los valores para las personas de nuestra población aparecen en la columna de la derecha.

μ_i: *Media poblacional del instrumento.* Es la media de todas las personas de la población para un instrumento determinado. Los valores para los instrumentos de nuestra población aparecen en la última fila.

Matemáticamente estos conceptos pueden expresarse del siguiente modo:

X_{pi}: Se considera una variable aleatoria.

$\mu_p = E_i(X_{pi})$: La puntuación universo de una persona es la esperanza matemática de sus puntuaciones empíricas a través de todos los instrumentos de medida. Conceptualmente sería equivalente a la puntuación verdadera en términos de la teoría clásica.

$\mu_i = E_p(X_{pi})$: La media poblacional del instrumento *i* es la esperanza matemática de las puntuaciones de todos los sujetos de la población en ese instrumento.

$\mu = E_p E_i(X_{pi})$: La gran media es la esperanza matemática de las puntuaciones de todas las personas de la población en todos los instrumentos de la población de instrumentos.

La puntuación empírica de una persona (X_{pi}) puede expresarse en función de los términos anteriores más un error de medida:

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + e_{pi}$$

Si se pasa μ al primer miembro de la igualdad:

$$X_{pi} - \mu = (\mu_p - \mu) + (\mu_i - \mu) + e_{pi}$$

Lo cual indica que lo que separa la puntuación de un sujeto de la gran media puede provenir de tres fuentes:

— $(\mu_p - \mu)$: Efecto debido a la persona.

— $(\mu_i - \mu)$: Efecto debido al instrumento.

— e_{pi}: Error residual.

Ahora bien, este idílico ejemplo, como ya se indicó, se aleja bastante de la mucho más espinosa realidad. ¿Qué es lo que normalmente nos vamos a encontrar en esa realidad? En primer lugar, nunca vamos a tener una población completa de instrumentos de medida ni de personas, tendremos una muestra. De modo que empíricamente nunca tendremos los valores reales de μ, μ_p y μ_i. Tendremos los que obtengamos en la muestra utilizada, que seguramente no coincidirán con los reales, pero, aunque no coincidan exactamente, ¿lo hacen razonablemente bien? En otras palabras, el valor de la muestra ¿es generalizable a la población? Ese es justo el problema que trata de resolver la TG: en qué medida una muestra de mediciones es generalizable a lo que se obtendría si, como en nuestro cándido ejemplo, se utilizase toda la población. Ese es el problema esencial, el cogollo de la TG.

El esquema del ejemplo presentado, el diseño, es uno de los muchísimos que se habrían podido establecer; se pueden llevar a cabo esquemas complejísimos en los que intervengan numerosos aspectos, no solo evaluadores como en nuestro caso, combinados de maneras varias. Pero en esencia, por más vueltas que se dé al diseño, el objetivo siempre es el mismo, a saber, comprobar en qué medida las condiciones bajo las que se realizó la medición generan una medida que representa lo que se hubiera obtenido si se hubiera trabajado con las poblaciones.

Para poder estimar los componentes del error, lo primero que ha de hacerse es *diseñar* una situación de medición en la que estén contempladas las posibles fuentes de error, es decir, se parte de un diseño determinado de recogida de datos, y lo que la TG hará, valiéndose del análisis de varianza, será estimar la cuantía de las distintas fuentes de error contempladas en el diseño de medición. En consecuencia, la tecnología operativa de la TG para estimar la fiabilidad de las mediciones descansa sobre dos pilares básicos:

a) El análisis de los distintos diseños que se pueden plantear.

b) Los análisis de varianza implicados en los diseños.

Otros conceptos importantes

— Objeto de la medición

En general, el objeto de la medición son las unidades medidas, sea cual sea su naturaleza, pero en el ámbito de las ciencias sociales los objetivos más habituales de la medición son las personas. En el ejemplo presentado en la tabla 2.4 el objeto de la medición serían los 10 aspirantes a trabajar. No obstante, cabe pensar en otros posibles destinatarios de la medición, tales como aulas, colegios, distritos u otros entes que interese evaluar.

— Faceta

Se da esa denominación a los aspectos o factores contemplados en el diseño de la medición. En nuestro ejemplo solo se consideró una faceta: los evaluadores. Cuantas más facetas se tienen en cuenta, más se complejiza el diseño; imagínese que en el ejemplo anterior además de los evaluadores se tuviesen en cuenta otras facetas tales como tipo de entrevista, sexo de los evaluadores u ocasiones de evaluación. No hay ninguna regla automática que le diga al investigador o profesional qué facetas debe incluir en su diseño. Han de tenerse en cuenta aquellas que se consideren relevantes para lo que se está midiendo, es decir, aquellas facetas que puedan afectar a las puntuaciones de las persoans y constituyan, por tanto, fuentes potenciales de error.

— Universo de observaciones admisibles

Esta expresión hace referencia a todas las puntuaciones de las personas que se consideran admisibles dentro del diseño planteado. En el ejemplo el universo de observaciones admisibles lo constituirían las puntuaciones de las personas llevadas a cabo por los evaluadores. En esta situación no sería admisible lógicamente una puntuación de una persona proveniente de un test. Una vez más este universo está en función del diseño planteado. Si además de los evaluadores hubiéramos tenido en cuenta distintos tipos de entrevistas, el universo serían las puntuaciones de las personas para cualquier evaluador y tipo de entrevista. En términos generales no es otra cosa que la población definida por el investigador al plantear el diseño de medida. Nótese que de los aspectos que no se incluyan en el diseño no se va a poder estimar su incidencia en la medida; por tanto, una recomendación genérica es plantear un diseño lo más general posible, que conlleve un universo de observaciones admisibles amplio. Claro que, operativamente, este noble consejo resulta un poco huero, pues finalmente será el profesional o el investigador quien tenga que decidir en cada caso concreto, en función de las circunstancias concurrentes.

— Universo de generalización

Se refiere a aquel aspecto, factor o faceta en el que el investigador está interesado en comprobar si los datos son generalizables. En nuestro ejemplo el único universo de generalización posible son los evaluadores. El investigador estaría interesado en comprobar si las medidas asignadas por una muestra de evaluadores a una muestra de personas son generalizables a la población de evaluadores. El universo de generalización son los evaluadores. Otros diseños más complejos permiten establecer otros universos de generalización de interés, por ejemplo, ocasiones de evaluación, tipos de entrevistas, etc. Un mismo diseño inicial va a permitir establecer más de un universo de generalización.

— Coeficiente de generalizabilidad

Viene a ser el equivalente al coeficiente de fiabilidad en la teoría clásica, e indica el grado en que una cierta medición es generalizable a la población de mediciones contempladas, es decir, en qué medida una muestra de mediciones representa al universo de generalización. Análogamente a la teoría clásica, sería el cociente entre la varianza universo y la empírica. Nótese que en función del universo de generalizabilidad elegido por el investigador va a poder existir más de un posible coeficiente G. En el ejemplo de los evaluadores, el coeficiente de generalizabilidad indicará en qué medida las calificaciones de una muestra de evaluadores son generalizables a la población de evaluadores. El coeficiente de generalizabilidad no va a ser la única forma que la TG tenga de expresar la cuantía de los errores de medida; también se pueden analizar los distintos componentes de la varianza generados por los ANOVA. Aunque todos los caminos lleven a Roma, en la literatura psi-

cométrica, seguramente influida por la lógica clásica del coeficiente de fiabilidad, se utiliza con más frecuencia el coeficiente de generallizabilidad.

8.3. Diseños de recogida de datos

Tras tanta niebla terminológica, no se olvide que el objetivo central de la TG es estimar la cuantía del error que afecta a las puntuaciones en función de las fuentes de las que provenga. Ahora bien, para poder estudiar esas fuentes de error hay que diseñar la recogida de datos (la medición) de tal forma que luego se puedan calcular los errores debidos a los distintos aspectos (facetas) contemplados. Por tanto, lo primero que hay que hacer antes de medir es decidir cómo se va a hacer, y en especial qué fuentes potenciales de error se van a contemplar. En nuestro sencillo ejemplo solo se tuvo en cuenta una faceta: los evaluadores. Si considerásemos que además son relevantes para la evaluación de los aspirantes el sexo de los evaluadores y el tipo de entrevista, tendríamos que tenerlo en cuenta y, por ejemplo, plantear un diseño de medida en el que se contemplasen esas dos facetas. Los posibles diseños son prácticamente ilimitados, están en función de lo que el investigador considere relevante para la medición que realiza. Lo que hay que tener claro es que de lo que no se incluya en el diseño no se podrá estimar su incidencia en la medición, y entrará a formar parte de esa caja negra inexcrutable que es el error aleatorio.

Una vez planteado el diseño, la obtención de los datos básicos para calcular los coeficientes de generalizabilidad y otros indicadores, se lleva a cabo mediante el análisis de varianza.

Más que plantear aquí una retahíla de diseños y sus correspondientes análisis, se van a ilustrar los cálculos fundamentales mediante el uso de dos diseños clásicos muy habituales en psicología y educación. Para diseños más complicados puede acudirse a los textos más exhaustivos citados al principio. Nótese, por ejemplo, que un libro tan clásico y recomendable como el de Kirk (1995) presenta 30 diseños distintos de análisis de varianza, sin pretensiones de exhaustividad. Por tanto, lo más importante es captar la lógica que subyace a todos ellos, para, llegado el caso, poder aplicarla al diseño que se tenga.

Diseños de una sola faceta

Se van a ejemplificar los cálculos principales de la TG utilizando un diseño, a la vez que sencillo, muy común en psicología y educación, con una sola faceta. De hecho, vamos a utilizar un ejemplo análogo al empleado para introducir los conceptos básicos, en el que la única faceta son los evaluadores. La cuestión básica a responder es en qué medida las calificaciones de los evaluadores son generalizables a la población de evaluadores, es decir, al universo de generalización formado por todos los evaluadores posibles. O, desde otro punto de vista equivalente, en qué medida las puntuaciones empíricas de las personas coinciden con sus puntuaciones universo.

Ejemplo

Sea una muestra aleatoria de 10 personas que son entrevistadas por una muestra también aleatoria de cuatro evaluadores para acceder a un trabajo. Las puntuaciones obtenidas por la muestra en una escala de 0 a 7 puntos aparecen en la tabla 2.5.

El coeficiente G es el cociente entre la varianza universo y la empírica, de modo que el problema consiste en hallar esos dos valores. Y es aquí donde entra en acción toda la maquinaria del análisis de varianza. Trabajar en el marco de la TG es hacerlo a base de exprimir los datos obtenidos al analizar los datos mediante ANOVA, de modo que el mejor

TABLA 2.5

		Evaluadores				
		E_1	E_2	E_3	E_4	$\bar{X}_p$
Personas	*a*	4	7	4	6	5,25
	b	4	4	2	3	3,25
	c	3	1	2	4	2,50
	d	5	7	5	4	5,25
	e	4	3	2	2	2,75
	f	0	3	3	4	2,50
	g	4	5	3	3	3,75
	h	4	5	5	4	4,50
	i	2	4	3	4	3,25
	j	6	7	5	4	5,50
	$\bar{X}_i$	3,6	4,6	3,4	3,8	3,85

consejo para aquellos investigadores o profesionales interesados en la TG es que repasen sus conocimientos del análisis de varianza, porque, a medida que se complejizan los diseños y se introducen más facetas, o esquemas de datos anidados, más se complica la ejecución de los cálculos. El paquete informático SPSS proporciona los datos necesarios en la mayoría de los diseños, y el programa GENOVA de Brennan y sus colaboradores, de la Universidad de Iowa, está específicamente diseñado para llevar a cabo los cálculos de la TG.

Si mediante alguno de los programas informáticos citados u otro cualquiera se analizan los datos de la tabla 2.5, que en términos de ANOVA corresponden a un diseño de medidas repetidas, puesto que, efectivamente, las personas son evaluados repetidamente por cuatro evaluadores, se obtiene una tabla como la 2.6.

TABLA 2.6

Fuentes de variación	Suma de cuadrados	Grados de libertad	Medias cuadráticas	Varianza estimada	% *V*
Personas (*p*)	50,60	9	5,622	1,089	43,46
Evaluadores (*i*)	8,30	3	2,767	0,150	5,98
Residual (*r*)	34,20	27	1,267	1,267	50,56

Veamos cómo se obtienen los datos de la tabla:

Suma de cuadrados total:

$$(SC_T) = \sum (X_{pi} - \bar{X}_T)^2 = (4 - 3{,}85)^2 + (7 - 3{,}85)^2 + \\ + (4 - 3{,}85)^2 + (6 - 3{,}85)^2 + (4 - 3{,}85)^2 + \cdots \\ + (5 - 3{,}85)^2 + (4 - 3{,}85)^2 = 93{,}10$$

Suma de cuadrados correspondiente a las personas:

$$SC_p = n_i \sum (\bar{X}_p - X_T)^2 = 4[(5{,}25 - 3{,}85)^2 + \\ + (3{,}25 - 3{,}85)^2 + \cdots + (3{,}25 - 3{,}85)^2 + \\ + (5{,}50 - 3{,}85)^2] = 50{,}60$$

Suma de cuadrados correspondiente a los evaluadores:

$$SC_i = n_p L(\bar{X}_i - X_T)^2 = 10[(3{,}6 - 3{,}85)^2 + \\ + (4{,}6 - 3{,}85)^2 + (3{,}4 - 3{,}85)^2 + (3{,}8 - 3{,}85)^2] = \\ = 8{,}30$$

Suma de cuadrados residual:

$$SC_r = SC_T - SC_p - SC_i = 93{,}10 - 50{,}60 - 8{,}30 = 34{,}20$$

Ya tenemos, por tanto, los datos de la primera columna de la tabla 2.6. La segunda columna son los grados de libertad, que vienen dados respectivamente por:

Personas	$p - 1 = 10 - 1 = 9$
Evaluadores	$i - 1 = 4 - 1 = 3$
Residual	$(p - 1)(i - 1) = (10 - 1)(4 - 1) = 27$

La tercera columna correspondiente a las medias cuadráticas se obtiene dividiendo las sumas cuadráticas por los correspondientes grados de libertad:

Personas	50,60/9 = 5,622
Evaluadores	8,30/3 = 2,767
Residual	34,20/27 = 1,267

A través de las esperanzas matemáticas de las medias cuadráticas se obtienen las estimaciones de las varianzas de los componentes, que vienen dadas por:

$$E(MC_p) = \sigma_e^2 + n_i\sigma_p^2$$
$$E(MC_i) = \sigma_e^2 + n_p\sigma_i^2$$
$$E(MC_r) = \sigma_e^2$$

Empezando por el final, puesto que $E(MC_r) = \sigma_e^2$, tenemos directamente el valor de $\sigma_e^2 = 1{,}267$. (Es-

trictamente no son los valores poblacionales, sino estimaciones de estos.)

Para calcular la varianza correspondiente a los evaluadores sustituimos en la segunda ecuación:

$$2{,}767 = 1{,}267 + 10\sigma_i^2$$

despejando:

$$\sigma_i^2 = \frac{2{,}767 - 1{,}267}{10} = 0{,}15$$

Utilizando la primera ecuación para la varianza de las personas:

$$5{,}622 = 1{,}267 + 4\sigma_p^2$$

despejando:

$$\sigma_p^2 = \frac{5{,}622 - 1{,}267}{4} = 1{,}089$$

Los cálculos anteriores pueden expresarse de forma compacta en función de las medias cuadráticas:

$$\sigma_p^2 = \frac{MC_p - MC_r}{n_i} = \frac{5{,}622 - 1{,}267}{4} = 1{,}089$$

$$\sigma_i^2 = \frac{MC_i - MC_r}{n_p} = \frac{2{,}767 - 1{,}267}{10} = 0{,}15$$

$$\sigma_e^2 = MC_r = 1{,}267$$

Finalmente, la última columna correspondiente al porcentaje de varianza se obtiene dividiendo cada valor de la columna anterior de las varianzas entre la suma de las tres y multiplicando por 100; esta columna representa, por tanto, el porcentaje de varianza correspondiente a cada fuente de variación. Basándose en los datos de esta columna, pueden hacerse todo tipo de interpretaciones acerca de los errores de medida, pues contiene los datos básicos sobre los que se van a hacer los cálculos posteriores del coeficiente G, como ahora se verá en el siguiente apartado.

Los datos de la tabla ya indican que la varianza residual es proporcionalmente muy elevada (50,56%), lo cual quiere decir que o bien:

a) Existe una fuerte interacción (*pxi*) entre personas y evaluadores, es decir, los evaluadores no actúan uniformemente con todas las personas.
b) O existe mucho error aleatorio de origen ignoto, no controlado en el diseño.
c) O ambas cosas a la vez, pues la varianza residual la componen las interacciones (*pxi*) y el error aleatorio.

Si se plantease en la práctica un caso semejante, habría que indagar qué ocurre con la interacción, dado que los evaluadores *per se* no introducen mucho error (5,98%) en la generalización. Veamos ahora cómo esto mismo queda reflejado de una forma global en el coeficiente G.

8.4. Coeficiente de generalizabilidad

A partir de la tabla 2.6 ya se puede calcular el coeficiente G, que viene dado por el cociente entre la varianza universo (σ_p^2) y la varianza empírica ($\sigma_p^2 + \sigma_e^2$):

$$\rho_g^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2} \qquad [2.39]$$

Aplicado a los datos de la tabla 2.6:

$$\rho_g^2 = \frac{1{,}089}{1{,}089 + 1{,}267} = 0{,}46$$

Dado que los valores posibles de ρ_g^2 están entre 0 y 1, este valor de 0,46 obtenido indica que las calificaciones dadas por los evaluadores son escasamente generalizables, es decir, lo que un evaluador asigna a las personas no es lo que estas obtendrían si fuesen evaluadas por todos los evaluadores de la población. Todo parece indicar que los evaluadores introducen un error notable en las calificaciones de las personas, por lo que sus puntuaciones están lejos de sus puntuaciones universo, y el error de medida es notable.

Más específicamente, este grado de generalizabilidad es el que correspondería a las calificaciones dadas a las personas por un evaluador cualquiera de la población; estrictamente no se refiere a uno de los evaluadores concretos utilizados en la muestra. De hecho, pueden calcularse coeficientes G específicos para cada uno de los evaluadores concretos utilizados en la muestra. Aunque esto puede tener interés en alguna situación, lo realmente importante es saber el grado de generalizabilidad que tienen las evaluaciones hechas por un evaluador cualquiera de la población.

A la vista de la escasa generalizabilidad, cabría plantearse en qué grado mejoraría esta si en vez de utilizar las calificaciones de un solo evaluador se promediasen las de varios. La fórmula para estos valores del coeficiente G es la siguiente:

$$\rho_G^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{n}} \qquad [2.40]$$

donde n es el número de evaluadores que se promedian.

Ejemplo

Si se promedian 2:

$$\rho_G^2 = \frac{1.089}{1.089 + \frac{1.267}{2}} = 0{,}63$$

El coeficiente G sube de 0,46 a 0,63.
Si $n = 3$:

$$\rho_G^2 = \frac{1.089}{1.089 + \frac{1.267}{3}} = 0{,}72$$

Con $n = 4$:

$$\rho_G^2 = \frac{1.089}{1.089 + \frac{1.267}{4}} = 0{,}77$$

Para $n = 5$:

$$\rho_G^2 = \frac{1.089}{1.089 + \frac{1.267}{5}} = 0{,}81$$

A medida que se ponderan las puntuaciones de más jueces, como es natural, la generalizabilidad de las puntuaciones aumenta, la medida se hace más fiable. Cuando n tiende a infinito, el coeficiente G tiende a uno.

Como probablemente ya habrá advertido el lector, el aumento del coeficiente al ir aumentando el número de jueces que se promedian viene dado por la fórmula de Spearman-Brown vista en la teoría clásica:

$$\rho_G^2 = \frac{n\rho_g^2}{1 + (n-1)\rho_g^2} \qquad [2.41]$$

donde:

- n: Número de jueces o instrumentos promediados.
- ρ_g^2: Coeficiente de generalizabilidad para un solo instrumento.
- ρ_G^2: Coeficiente de generalizabilidad tras promediar los instrumentos.

Por ejemplo, en el caso de $n = 3$:

$$\rho_G^2 = \frac{3(0{,}46)}{1 + (3-1)0{,}46} = 0{,}72$$

que coincide con el valor obtenido mediante la fórmula [2.40].

Cálculo del coeficiente de generalizabilidad en función de las medias cuadráticas

Para los cálculos del coeficiente G resulta más sencillo utilizar los valores de las medias cuadráticas, sin necesidad de calcular las varianzas; la fórmula viene dada por:

$$\rho_g^2 = \frac{MC_p - MC_r}{MC_p + (n_i - 1)MC_r} \qquad [2.42]$$

Para los datos de la tabla 2.6:

$$\rho_g^2 = \frac{5.622 - 1.267}{5.622 + (4-1)1.267} = 0{,}46$$

Esta fórmula resulta mucho más directa operativamente, pero no es tan clara para captar conceptualmente el significado del coeficiente G como la [2.39].

A veces puede ocurrir que al estimar las varianzas estas tengan signo negativo, lo cual no es teóricamente posible, puesto que las varianzas, al provenir de datos elevados al cuadrado, han de ser necesariamente positivas. Cuando esto ocurre, la solución pragmática más habitual es igualar a cero la varianza negativa, si bien los especialistas que han analizado este paradójico asunto han sugerido algunas alternativas más sofisticadas (Brennan, 1983, 2001; Cronbach et al., 1972; Shavelson, Webb y Rowley, 1989).

Otros diseños de una faceta

El diseño utilizado para ilustrar los cálculos se denomina cruzado (*pxi*); todos los evaluadores evalúan a todas las personas, los evaluadores están cruzados con las personas. Pero son posibles otros diseños en los que no ocurra esto; por ejemplo, podría darse el caso de que varios evaluadores juzgasen a cada persona, siendo distintos los evaluadores para cada persona. Incluso un caso particular del anterior es que cada persona fuese evaluada por un solo evaluador, distinto para cada persona. Este tipo de diseños se denominan «anidados»; en los dos casos anteriores los evaluadores se hallarían anidados en las personas (*i*:*p*).

El coeficiente de generalizabilidad para estos diseños viene dado por:

$$\rho_g^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_i^2 + \sigma_e^2}{n}} \qquad [2.43]$$

donde *n* es el número de evaluadores que se promedia; cuando se trata de uno solo, la fórmula se transforma en:

$$\rho_g^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_i^2 + \sigma_e^2} \qquad [2.44]$$

8.5. Estudios de generalizabilidad y estudios de decisión

La TG hace una distinción conceptual entre los así llamados estudios de generalizabilidad (estudios-G) y los estudios de decisión (estudios-D). Los estudios-G son los planteamientos genéricos del diseño, a través del cual, y valiéndose del ANOVA, se estiman las varianzas de los distintos componentes. Son, por así decirlo, el marco general en el que están desplegados todos los datos que posteriormente se utilizarán por el investigador para resolver un problema concreto. A estos problemas concretos, sean los que sean, que utilizan los datos de los estudios-G se les denomina estudios-D. Un estudio-G puede servir de marco básico para distintos estudios-D. Por ejemplo, los datos de la tabla constituyen un estudio-G, mientras que la decisión de calcular el coeficiente de generalizabilidad (ρ_g^2) para tres evaluadores, o para cinco, o los que sean, serían estudios-D.

Cuanto más amplio sea el estudio-G, más posibilidades abrirá de plantear distintos estudios-D. Debido a esta posibilidad de plantear diversos estudios-D, la TG se vuelve compleja, pues en cada caso serán distintos los datos del estudio-G a utilizar, y lo que en cada caso se considere varianza universo y varianza error variará.

Facetas fijas y aleatorias

Otro factor que introduce complejidad en los cálculos de la TG es que las facetas utilizadas sean fijas o aleatorias. Se considera que una faceta es fija cuando los valores considerados agotan la población, mientras que en el caso de las aleatorias se asume que los niveles de la faceta constituyen una muestra aleatoria. Según se trate de uno u otro caso, los componentes para los cálculos serán distintos.

Lo más habitual es que las facetas se consideren aleatorias, pues se ajusta más a la mayoría de los problemas con los que se enfrentan los investigadores y profesionales, al menos en el campo de las ciencias sociales.

En el ejemplo de la tabla 2.5 la faceta «evaluadores» se asumió aleatoria: los cuatro evaluadores utilizados se consideran una muestra aleatoria de la población de evaluadores; por tanto, las fórmulas

propuestas para los cálculos realizados son las pertinentes para el caso de facetas aleatorias. Una gama variada de posibilidades con sus correspondientes formulaciones puede consultarse en Brennan (1983, 2001), Crocker y Algina (1986) o en Shavelson y Webb (1991).

8.6. Error típico de medida

En la teoría clásica, una forma de expresar los errores de medida, además del coeficiente de fiabilidad, era el error típico de medida, utilizado sobre todo para establecer intervalos de confianza en torno a las puntaciones empíricas para así estimar las verdaderas. Con la TG ocurre lo mismo, y el error típico de medida para el caso de un diseño cruzado (*pxi*) viene dado por:

$$\sigma_{e(g)} = \sqrt{\sigma_i^2 + \sigma_e^2} \qquad [2.45]$$

Si se promedian los evaluadores, la varianza error se divide entre n, el número de evaluadores promediados, y el error típico de medida viene dado entonces por:

$$\sigma_{e(g)} = \sqrt{\frac{\sigma_i^2 + \sigma_e^2}{n}} \qquad [2.46]$$

Ejemplo

Un evaluador (tabla 2.4) asignó a una persona una puntuación de 6. Al nivel de confianza del 95%, ¿qué puntuación universo se estima que obtendrá dicha persona?

Análogamente a como se procedía en la teoría clásica:

1. NC 95%: $Z_c = \pm 1{,}96$.
2. $\sigma_{e(g)} = \sqrt{\sigma_i^2 + \sigma_e^2} = \sqrt{0{,}150 + 1{,}267} = 1{,}190$.
3. Error máximo: $(Z_c)(\sigma_{e(g)}) = 1{,}96 \times 1{,}190 = 2{,}332$.
4. Intervalo confidencial: $X_{pi} \pm$ E. máximo = $= 6 \pm 2{,}332$.

$$3{,}668 \leqslant \mu_p \leqslant 8{,}332$$

Obsérvese la coherencia de esta estimación con la obtenida mediante el coeficiente G. El intervalo confidencial resulta muy amplio; no se puede estimar la puntuación de la persona con precisión, hay mucho error en la medida. Recuérdese que el coeficiente G era bajo (0,46).

Si la calificación de la persona anterior que obtuvo 6 puntos proviniese de haber promediado las calificaciones de tres jueces, tendría que ser más fiable, y el intervalo confidencial debería ser más estrecho. Veámoslo:

1. NC 95%: $Z_c = \pm 1{,}96$.
2. $\sigma_{e(g)} = \sqrt{\frac{\sigma_i^2 + \sigma_e^2}{n}} = \sqrt{\frac{0{,}150 + 1{,}267}{3}} = 0{,}687$.
3. Error máximo: $(Z_c)(\sigma_{e(g)}) = 1{,}96 \times 0{,}687 = 1{,}346$.
4. Intervalo confidencial: $X_{pi} \pm$ E. máximo = $= 6 \pm 1{,}346$.

$$4{,}654 \leqslant \mu_p \leqslant 7{,}346$$

Efectivamente, la amplitud del intervalo es ahora notablemente inferior. Teóricamente llegaría a ser nula si no hubiese errores de medida.

Estos errores típicos de medida se aplican también en el caso de los diseños anidados.

Decisiones absolutas y decisiones relativas

Hay que advertir de nuevo que las fórmulas del error típico de medida varían en función del tipo de estudios planteados; por tanto, las aquí presentadas no son válidas para otros tipos de diseños. Además, las fórmulas [2.43] y [2.44] son pertinentes para en lo que la TG se denominan *decisiones absolutas,* pero no lo son para *decisiones relativas.*

Se entiende por decisiones absolutas aquellas que se toman sobre las puntuaciones de las personas sin tener en cuenta al resto de las personas, por ejemplo, establecer un intervalo confidencial en torno a su puntuación empírica para estimar a un cierto nivel de confianza su puntuación universo, como se ha hecho en el ejemplo anterior.

Decisiones relativas serían aquellas en las que se tiene en cuenta la posición relativa de las personas,

como cuando se comparan unas personas con otras y se estudian las diferencias entre sus puntuaciones, o se toman decisiones basadas en la posición relativa de las personas en el grupo.

8.7. Diseños de dos facetas

Se habla de un diseño de dos facetas cuando son dos los aspectos o facetas tenidos en cuenta en el esquema con el que se realiza la medición. Por ejemplo, una muestra de 10 personas son evaluadas por tres jueces, tanto por escrito como oralmente. En este diseño se contemplan dos facetas:

- La faceta *evaluadores,* con tres niveles correspondientes a cada uno de los evaluadores.
- La faceta *modalidad,* con dos niveles: oral, escrito.

Este tipo de esquemas de evaluación se da con cierta frecuencia en muchos campos de las ciencias sociales y de la salud, y la cuestión básica que se plantea aquí es si las puntuaciones de las personas son generalizables o, por el contrario, están muy mediatizadas por la modalidad evaluativa y/o evaluador.

Supongamos que se evalúa a 10 personas con el diseño anterior, obteniéndose tras el correspondiente análisis de varianza los datos de la tabla 2.7.

TABLA 2.7

Fuentes de variación	Suma de cuadrados	Grados de libertad	Medias cuadráticas	Varianza estimada	% *V*
Personas (*p*)	70,54	9	7,84	0,462	12,8
Evaluador (*i*)	10,70	2	5,35	0,026	0,7
Modalidad (*m*)	9,82	1	9,82	0,194	5,4
PersxEva. (*pi*)	62,50	9	3,47	1,210	33,5
PersxOca. (*pm*)	23,85	18	2,65	0,533	14,7
EvaxOca. (*im*)	4,80	2	2,40	0,135	3,7
Resid. (*pim* + *e*)	18,90	18	1,05	1,050	29,1

Nota: Para los cálculos de las varianzas estimadas, véase apéndice (2.40).

Coeficiente de generalizabilidad

La fórmula del coeficiente *G,* como en el caso de una faceta, viene dada por el cociente entre la varianza universo y la varianza observada:

$$\rho_g^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pm}^2}{n_m} + \frac{\sigma_e^2}{n_i n_m}} \qquad [2.47]$$

donde:

- n_i: Es el número de evaluadores que se consideren en el estudio *D.*
- n_m: Es el número de modalidades del estudio *D* que se plantee.

Sustituyendo en la fórmula los valores de la tabla 2.7:

$$\rho_g^2 = \frac{0{,}462}{0{,}462 + \frac{1{,}210}{3} + \frac{0{,}533}{2} + \frac{1{,}050}{(3)(2)}} = 0{,}35$$

Dado que la generalizabilidad resulta tan baja, el investigador puede considerar, por ejemplo, la posibilidad de utilizar más evaluadores, por ejemplo, seis en vez de tres. En ese caso, con $n_i = 6$, el coeficiente sería:

$$\rho_g^2 = \frac{0{,}462}{0{,}462 + \frac{1{,}210}{6} + \frac{0{,}533}{2} + \frac{1{,}050}{(6)(2)}} = 0{,}45$$

La mejora no es notoria, pasa de 0,35 a 0,45 cuando se duplican los evaluadores, por lo que el problema de la escasa generalizabilidad no parece provenir de los evaluadores. Aquí es cuando resulta muy aclaratorio acudir a las columnas de varianza estimada y ver cuáles son las fuentes que contribuyen más al error o, lo que es lo mismo, a esta escasa generalizabilidad. Efectivamente, las fuentes que más error aportan son la interacción personas × evaluadores (*pi*), con un 33,5%; la interacción personas × modalidad (*pm*), un 14,7%, y el residual que proviene de las interacciones personas × evaluador × modalidad, más el error aleatorio. La primera interacción (*pi*) indicaría que por alguna razón, que habrá que averiguar, los evaluadores no actúan igual con todas las personas. Análogamente, la interacción (*pm*) pone de manifiesto que las modalidades oral/escrito afectan diferencialmente a las personas, interactúan con ellas. Algo similar ocurre con las interacciones (*pim*): resultan demasiado elevadas. Nótese que todos estos términos están en el denominador del coeficiente G, forman parte de la varianza observada, de modo que cuanto mayores sean, menor será el coeficiente G. Esta partición de la varianza aplicada a la fiabilidad es la aportación central de la TG, que permite diseccionar e indagar el porqué de un coeficiente de fiabilidad bajo.

Conviene recordar que con esas dos mismas facetas cabe plantear otros muchos posibles diseños; aquí se ha utilizado lo que se suele denominar diseño cruzado ($p \times i \times m$), pues todas las personas son evaluadas en las dos modalidades por todos los evaluadores. Cuando no se produce este cruce, se habla de diseños anidados. Siguiendo con el mismo ejemplo, piénsese que por razones de economía los evaluadores se repartiesen y un grupo evaluase las pruebas orales y otro las escritas, en vez de que todos los evaluadores evaluasen todo, como ocurría en el diseño del ejemplo. En este caso se habla de un diseño anidado, los evaluadores se encontrarían anidados en la modalidad. Pues bien, si se utilizase este diseño, los datos para calcular el coeficiente G variarían, pues las varianzas universo y observada están en función del diseño que se plantee. Los textos más avanzados sobre TG suelen proporcionar los componentes para toda una gama de diseños (Brennan, 1983, 2001; Crocker y Algina, 1986; Shavelson y Webb, 1991), tanto de dos facetas como de más.

9. FIABILIDAD DE LOS TEST REFERIDOS AL CRITERIO

9.1. Definición

El objetivo central de los test tratados hasta ahora en este libro era ordenar o escalar a las personas según su puntuación en un determinado rasgo o variable psicológica, bien fuese de carácter cognoscitivo, como las aptitudes, de personalidad (extraversión, neuroticismo, etc.), actitudes o de cualquier otro tipo. Estos test se construyen con ese objetivo en mente y, por tanto, se intenta que discriminen al máximo entre las personas para así poder escalarlas con mayor precisión o fiabilidad. La puntuación de una persona se expresa a partir de ciertas normas establecidas en función de las puntuaciones del grupo. Por ejemplo, se dice que Palomino Molero ocupa el percentil 80, es decir, puntúa por encima del 80% de sus compañeros. La puntuación de una persona será alta o baja en función de su posición relativa en el grupo, posición que se puede expresar de diversos modos, como se verá en el apartado 9.5. Estos test suelen denominarse genéricamente test referidos a normas, o normativos, y son los más habituales en el ámbito de la psicología.

Pero a partir de los años cincuenta y sesenta, con el predominio del enfoque conductual en psicología y el aumento de los métodos de enseñanza programada y de las máquinas de enseñar, aparece la necesidad de construir test que evalúen directamente el conocimiento que tienen los estudiantes de los objetivos programados. Más que una variable psicológica o rasgo, estos test van dirigidos a evaluar un dominio o criterio concreto de interés; de ahí su denominación de test referidos al criterio (TRC). En ellos no se pone el énfasis en analizar las diferencias entre las personas, sino más bien se trata de ver en qué medida cada persona domina el criterio de interés previamente definido. No se trata de ubicar la posición relativa de la persona, sino de detectar en qué grado conoce los objetivos. Una forma habitual de expresarlo es mediante el porcentaje de cuestiones respondidas correctamente. Así, se dirá, por ejemplo, que Agapito Hito domina el 90% del criterio. Naturalmente, estos test referidos al criterio también terminan ordenando a las personas, según el dominio que estas tengan del criterio evaluado, pues unas dominarán un porcentaje mayor que

otras; pero ese no es el objetivo prioritario que guía su construcción, como lo hacía en el caso de los test normativos. En los TRC la discriminación máxima entre las personas no es una propiedad específicamente buscada. Hay que señalar que los TRC no solo encajan a la perfección en el ámbito de la evaluación educativa, en general su enfoque es apropiado para evaluar cualquier área de conocimiento.

El concepto y la propia denominación de los TRC tienen su origen en un magistral artículo de tres páginas publicado por Robert Glaser en 1963 en la revista *American Psychologist*. Actualmente la literatura sobre los TRC es abundantísima. Una excelente panorámica de su desarrollo en los últimos cuarenta años puede consultarse en Hambleton et al. (2016), y buenas exposiciones pueden consultarse en Berk (1984a), Hambleton (1980), Hambleton et al. (1978), Popham (1978) o Shrock y Coscarelli (2007), entre otros. Por su parte, Nitko (1984) lleva a cabo un análisis minucioso de la definición y concepto de los TRC.

Pensarán muchos lectores que en realidad esto de los TRC no es nada novedoso, que siempre ha habido exámenes y test cuya finalidad era evaluar un dominio concreto de conocimientos o habilidades, y que ello se venía haciendo habitualmente en psicología educativa y del trabajo. Es cierto; lo que ocurre es que con el énfasis y sistematización surgidos a partir de los trabajos pioneros (Glaser, 1963; Glaser y Klaus, 1962; Popham y Husek, 1969) se va a desarrollar todo un refinamiento técnico y psicométrico para la elaboración y análisis de este tipo de test, ya que la metodología clásica al uso no se ajustaba bien a los nuevos planteamientos. De este modo, los test referidos al criterio han propiciado el desarrollo de ciertos ámbitos de la medición psicológica y educativa implicados en su desarrollo y aplicación. Hambleton (1994a) cita seis campos principales impulsados por los TRC.

1. Un primer efecto muy positivo ha sido el de obligar a profesores y constructores de test a definir con mayor claridad y operatividad los objetivos o criterios de interés, en la línea de la evaluación conductual, para así poder construir los test correspondientes para su evaluación.
2. Obligan a muestrear exhaustivamente los objetivos a evaluar y exigen sumo cuidado a la hora de confeccionar los ítems. No en vano, a raíz del impulso de los TRC florece toda una tecnología para la escritura de los ítems (Roid, 1984; Roid y Haladyna, 1980, 1982), apareciendo además un variado abanico de formatos alternativos al omnipresente de elección múltiple.
3. Se potenciaron nuevas formas para evaluar la fiabilidad y validez de los test, como se verá en el siguiente apartado, pues las utilizadas para los test referidos a normas no siempre resultaban las más apropiadas.
4. Dado que con gran frecuencia el uso de los TRC exigía dividir a las personas en dos grupos, las que dominaban el criterio y las que no, se desarrolló toda una tecnología psicométrica para establecer de un modo adecuado los puntos de corte para determinar quién pasa y quién falla.
5. Los TRC, al centrarse operativamente en los objetivos específicos, han sido altamente beneficiosos para el diagnóstico de las deficiencias de aprendizaje. Permiten detectar los puntos fuertes y débiles de las personas y ayudar a los profesores a tomar decisiones sobre la enseñanza. Además, fomentan que los profesores hagan más hincapié en el dominio que los estudiantes tienen de la materia que en el mero análisis de las diferencias entre ellos.
6. Finalmente, ha hecho que los profesores adquieran conocimientos en el campo de la evaluación de los estudiantes. Esto es de suma importancia, pues con demasiada frecuencia algo tan relevante como la evaluación adecuada y rigurosa de los estudiantes se deja al sentido común de los profesores, poniendo en peligro la obligada equidad evaluativa. Por otras latitudes no se tiene tanta confianza en la ciencia infusa de los profesores en materia de medición educativa; por ejemplo, los dos grandes sindicatos de profesores americanos han publicado unos estándares técnicos (American Federation of Teachers, NCME y NEA, 1990) sobre la competencia de los profesores para evaluar a los alumnos, en los que los TRC desempeñan un papel central. Esperemos que en otros países cunda el ejemplo, pues antes que nada está el derecho del estu-

> diante a una evaluación justa, condición *sine qua non* para una enseñanza rigurosa y de calidad. Parafraseando al gran físico y matemático Lord Kelvin, mal podemos mejorar la enseñanza si no empezamos por evaluarla con rigor, pues lo que no se mide no se puede mejorar.

En este apartado se exponen algunas de las técnicas para el cálculo de la fiabilidad de los test referidos al criterio, pues, como se irá viendo, los coeficientes de fiabilidad vistos hasta ahora no siempre son los más adecuados para utilizar con los TRC.

9.2. Métodos de estimación de la fiabilidad

El problema de la fiabilidad de los test referidos al criterio en esencia es el mismo que el de los test clásicos referidos a las normas. En ambos casos se trata de estimar el grado de error incrustado en las mediciones. Puesto que los dominios o criterios de interés a evaluar con los TRC suelen ser amplios, el test utilizado para hacerlo es una de las posibles muestras de ítems. Si el test fuese completamente fiable, el porcentaje de ítems contestado correctamente por cada persona coincidiría con el porcentaje que estas obtendrían si se utilizase el dominio completo. En líneas generales, la fiabilidad trata de estimar en qué medida ambos porcentajes coinciden. Naturalmente, el porcentaje de aciertos en el dominio no se puede obtener empíricamente, por lo que para estimar la fiabilidad habrá que recurrir a procedimientos indirectos, como ocurría en la teoría clásica.

Los métodos clásicos vistos hasta ahora, tales como test-retest, formas paralelas o consistencia interna (coeficiente alfa), pueden utilizarse como una primera aproximación al cálculo de la fiabilidad de los TRC, pero en este apartado se van a proponer otros acercamientos más específicos. Se preguntará el lector por qué esos métodos clásicos no son del todo satisfactorios para estimar la fiabilidad de los TRC, que al fin y al cabo no son otra cosa que test clásicos destinados a evaluar un dominio específico de contenidos, sean estos de la naturaleza que sean. La razón fundamental es que en la práctica la mayoría de los TRC tienen como finalidad clasificar a las personas en dos categorías, las que dominan el criterio y las que no lo dominan, aunque nada impide hacer más categorías. De este modo, la fiabilidad toma los derroteros de evaluar la consistencia o precisión de estas clasificaciones, adoptando métodos relacionados con la toma de decisiones. Desde un punto de vista más teórico, los métodos clásicos de fiabilidad resultan óptimos cuando el test se ha construido pensando en que debe maximizar la discriminación entre las personas, que no es el caso de los TRC. Además, el concepto de test paralelos, piedra angular de la fiabilidad clásica, no representa un papel central en los TRC, que más bien constituyen en teoría muestras aleatorias de los contenidos del dominio. Por esa vía indirecta, si realmente fuesen muestras aleatorias, deberían ser paralelos en el sentido clásico. De hecho, como luego se verá, algunos de los métodos expuestos asumen el paralelismo clásico.

Las técnicas para evaluar la fiabilidad pueden clasificarse de distintas maneras, según se atienda a un criterio u otro. Por razones de claridad, aquí se han clasificado en dos grandes bloques: aquellas que exigen dos aplicaciones del test, bien sea del mismo test o de formas paralelas, y las que solo exigen una aplicación. Dentro de este segundo apartado, a su vez, se hacen tres subgrupos en función de cómo se utilice el punto de corte para las clasificaciones. Un tratamiento detallado de la fiabilidad de los TRC puede consultarse en el libro de Berk (1984a), especialmente los capítulos del propio Berk, Subkoviak y Brennan. Muy interesantes resultan el análisis de Hambleton y Slater (1997), y, a un nivel más introductorio, el libro de Crocker y Algina (1986). Una síntesis de la problemática y enfoques de la fiabilidad de los TRC puede verse en Han y Rudner (2016).

Formas paralelas

Se recogen aquí los coeficientes de fiabilidad cuando se dispone de dos aplicaciones del mismo test a una muestra de personas, o de la aplicación de dos formas paralelas del test. En esas circunstancias, si se establece un punto de corte y en cada test se clasifica a las personas en dos grupos, las que superan el punto de corte y las que no lo superan, si existiese una fiabilidad perfecta la clasificación resultante debería ser idéntica para ambos test. Pues bien, los coeficientes de fiabilidad que se van a ver en este apartado tratan de estimar en qué medida las clasi-

ficaciones hechas con un test coinciden con las hechas por otro, o por el mismo aplicado en dos ocasiones. Aunque la filosofía básica es la misma que la vista en la aproximación clásica, debido al uso habitual de los TRC para llevar a cabo clasificaciones, la forma operativa de calcular la fiabilidad varía, tratándose más bien de índices de acuerdo entre las clasificaciones. En la literatura psicométrica es muy habitual denominar *masters* a las personas que superan el punto de corte y *no masters* a las que no lo superan. Bien es verdad que a veces las clasificaciones no solo tienen dos categorías; por ejemplo, se puede clasificar a las personas según su rendimiento en bajos, medios, altos, etc. Los coeficientes pueden utilizarse con cualquier número de categorías.

Veamos un ejemplo sobre el que se ilustrarán los coeficientes. Sea una muestra de veinte personas a las que se aplicaron dos formas paralelas de un test de vocabulario de diez ítems. Se considera que para superar la prueba hay que sacar una puntuación igual o superior a 7. Las puntuaciones de las personas en ambas formas del test aparecen en la tabla 2.8.

Teniendo en cuenta que para superar la prueba hay que obtener 7 puntos o más, compruebe el lector que las 20 personas de la tabla 2.8 quedan clasificadas según superen o no superen las pruebas de acuerdo con la tabla 2.9. Una primera aproximación elemental a la fiabilidad sería ver si los porcentajes de personas que superan la prueba son los mismos en ambas formas del test. En nuestro caso, con la forma A superan el criterio siete de las 20 personas (35%), y con la forma B, ocho de las 20 (40%). La fiabilidad sería perfecta cuando los porcentajes fuesen los mismos. Este razonamiento, cuya introducción en el ámbito de los TRC suele atribuirse a Carver (1970), tiene un claro inconveniente que lo hace desaconsejable. Los porcentajes podrían coincidir, pero no ser las mismas personas las que superasen ambas pruebas, en cuyo caso el indicador conduce a un claro error, dando una falsa idea de fiabilidad donde no la hay. Ello se debe a que este indicador no tiene en cuenta la consistencia de las clasificaciones individuales. Los índices que se verán a continuación sí tienen en cuenta esta consistencia.

TABLA 2.8

Persona	Forma *A*	Forma *B*
1	4	7
2	7	8
3	5	4
4	6	6
5	8	9
6	6	4
7	5	7
8	3	2
9	9	4
10	3	2
11	7	10
12	5	3
13	4	4
14	10	9
15	3	2
16	3	4
17	8	7
18	8	7
19	2	3
20	0	1

Coeficiente p_o

Este coeficiente fue propuesto por Hambleton y Novick (1973), y posteriormente complementado por Swaminathan, Hambleton y Algina (1974). La idea es sencilla: trata de reflejar en qué medida las clasificaciones hechas a partir de ambos test coinciden. Si se observa la tabla 2.9, se ve que, salvo tres personas, las otras 17 son clasificadas del mismo modo por los dos test. Parece, por tanto, que la fiabilidad de la clasificación es elevada. El coeficiente p_o permite expresar esta fiabilidad por medio de la proporción de coincidencias observadas. Su fórmula viene dada por:

$$p_o = \frac{F_c}{N} \qquad [2.48]$$

TABLA 2.9

Forma A	Forma B		
	Superan	No superan	Total
Superan	6	1	7
No superan	2	11	13
Total	**8**	**12**	**20**

donde:

F_c: Es el número de personas (frecuencia) en las que ambos test coinciden en la clasificación.
N: Es el número total de personas.

Para los datos de la tabla 2.9:

$$p_o = \frac{6 + 11}{20} = 0{,}85$$

El valor máximo del coeficiente p_o es 1, que ocurriría cuando las clasificaciones hechas con las dos formas del test fuesen exactamente las mismas, es decir, cuando todas las frecuencias estuviesen en las casillas de la diagonal principal. El valor mínimo es el que cabe esperar por mero azar, y viene dado en función de las frecuencias marginales de la tabla. Para los datos de la tabla 2.9, las coincidencias por mero azar (p_a) se calcularían del siguiente modo:

$$\text{Casilla 1-1: } \frac{7 \times 8}{20} = 2{,}8$$

$$\text{Casilla 2-2: } \frac{13 \times 12}{20} = 7{,}8$$

$$p_a = \frac{2{,}8 + 7{,}8}{20} = 0{,}53$$

Como se puede observar, el uso de los test mejora considerablemente las clasificaciones que cabría esperar por mero azar, pasando de 0,53 a 0,85. Para una interpretación adecuada, siempre hay que tener en cuenta lo que cabe esperar por mero azar. Precisamente, el coeficiente *kappa* que vamos a ver a continuación contempla en su formulación los aciertos por azar.

Aunque se han utilizado solo dos categorías, superar y no superar, el coeficiente p_o puede calcularse análogamente para cualquier número de categorías.

Coeficiente kappa

Propuesto por Cohen en 1960, es uno de los coeficientes más populares y reseñados en la literatura psicométrica. Los primeros en aconsejar su utilización en el contexto de la fiabilidad de los TRC fueron Swaminathan, Hambleton y Algina (1974).

Su fórmula viene dada por:

$$K = \frac{F_c - F_a}{N - F_a} \qquad [2.49]$$

donde:

F_c: Frecuencia de coincidencia, o número de casos en los que las clasificaciones de ambos test coinciden.
F_a: Frecuencia de azar, o número de casos en que cabe esperar por mero azar que las clasificaciones de ambos test coincidan.
N: Número total de personas de la muestra.

Aplicación a los datos de la tabla 2.9.

En primer lugar, se calcula el valor de las frecuencias esperadas por azar:

$$\text{Casilla 1-1: } \frac{8 \times 7}{20} = 2{,}8$$

$$\text{Casilla 2-2: } \frac{12 \times 13}{20} = 7{,}8$$

$$F_a = 2{,}8 + 7{,}8 = 10{,}6$$

Las frecuencias de coincidencia vienen dadas por:

$$F_c = 6 + 11 = 17$$

Aplicando la fórmula:

$$K = \frac{17 - 10{,}6}{20 - 10{,}6} = 0{,}68$$

El valor máximo del coeficiente *kappa* es 1, cuando la fiabilidad es perfecta; pero el mínimo depende de las frecuencias marginales de la tabla. En el contexto de la fiabilidad los valores negativos no tienen sentido, los cercanos a cero indicarían que las clasificaciones hechas por los test no mejoran el azar. Brennan y Prediger (1981) hacen un buen análisis de las posibilidades y límites del coeficiente kappa.

La fórmula del coeficiente *kappa* puede expresarse en función de las proporciones en vez de las frecuencias:

$$K = \frac{P_c - P_a}{1 - P_a} \qquad [2.50]$$

Ni que decir tiene que el resultado obtenido con ambas fórmulas es el mismo, como puede comprobar el lector aplicando esta fórmula a los datos de la tabla 2.9. El valor del coeficiente *kappa* es muy similar al coeficiente de correlación de Pearson para datos dicotómicos, es decir, al coeficiente Φ.

Como ya se ha señalado, los dos coeficientes (p_o y K) pueden aplicarse cuando los test referidos al criterio se utilizan para clasificar a las personas en más de dos categorías. Como ejercicio, suponga el lector que en la tabla 2.8, en vez de dos categorías, se han hecho tres: baja (puntuaciones 0-3), media (puntaciones 4-7) y alta (puntuaciones 8-10). Elabore la tabla correspondiente y calcule los coeficientes p_o y K.

Significación estadística del coeficiente kappa

La significación estadística del coeficiente *kappa* puede someterse a prueba utilizando el error típico de medida propuesto por el propio Cohen (1960):

$$\sigma_e = \sqrt{\frac{F_a}{N(N - F_a)}}$$

Hagámoslo para los datos del ejemplo anterior al nivel de confianza del 95%.

La hipótesis nula y la alternativa serán:

$$H_0: \ K = 0$$
$$H_1: \ K \neq 0$$

Error típico de medida:

$$S_e = \sqrt{\frac{10{,}6}{20(20 - 10{,}6)}} = 0{,}24$$

Intervalo confidencial:

$$0{,}68 \pm (1{,}96)(0{,}24)$$
$$(0{,}21 \leqslant K \leqslant 1{,}00)$$

Dado que el valor $K = 0$ no está dentro del intervalo confidencial, se rechaza la hipótesis nula, y el coeficiente resulta estadísticamente significativo.

Ciertamente, el cálculo de la significación estadística de K en la investigación aplicada puede parecer algo trivial, pues generalmente cabe esperar que la fiabilidad sea considerable. Como señala Cohen (1960), tal vez pueda ser más útil de cara a establecer los mínimos exigibles en determinadas situaciones.

Fleiss, Cohen y Everitt (1969) proponen otra fórmula para el error típico de medida técnicamente más adecuada, pero mucho más compleja. Sin embargo, las diferencias empíricas entre ambas son mínimas. Además, dado que σ_e, según Cohen (1960), suele ser algo mayor que Fleiss et al. (1969), la prueba es más conservadora, lo cual nunca viene mal cuando de análisis de datos se trata. Hanley (1987) propuso una simplificación atinada para la formulación de Fleiss et al. (1969).

Finalmente, señalar que el coeficiente *kappa* fue extendido (Cohen, 1968) para situaciones en las que todos los desacuerdos no se consideran igual de importantes, dándose distintas ponderaciones a algunos de ellos según cierto criterio, y calculándose en estos casos el coeficiente *kappa* ponderado.

Una sola aplicación del test

En la mayoría de las situaciones aplicadas los profesionales no tienen la posibilidad de utilizar dos formas paralelas del test, ni van a poder aplicar el mismo test dos veces, de modo que tendrán que arreglarse con una sola aplicación. En ese caso los coeficientes de fiabilidad del apartado anterior no se pueden utilizar tal como se expusieron allí, pues exigían disponer de las puntuaciones en dos aplicaciones. En este apartado se exponen cinco métodos que solo exigen aplicar el test una vez, si bien son muy distintos entre sí, como se irá viendo. El primero (Subkoviak, 1976) y el segundo (Huynh, 1976) permiten estimar los coeficientes p_o y *kappa*, vistos en el apartado anterior. El tercero (Livingston, 1972) y el cuarto (Brennan y Kane, 1977) tienen en cuenta las distancias de las puntuaciones de las personas al punto de corte establecido para clasificarlos, y el quinto se basa en la teoría de la generalizabilidad.

Método de Subkoviak

El método propuesto por Subkoviak (1976) permite estimar el valor de los coeficientes p_o y *kappa* cuando solo se dispone de una aplicación del test. Basándose en determinadas asunciones, el método de Subkoviak simula las puntuaciones de una segunda forma paralela del test, de la cual se carece en la práctica. Procedimientos similares al de Subkoviak fueron propuestos por Huynh (1976) y Marshall y Haertel (1976). Una buena exposición y un análisis comparativo de los tres pueden consultarse en Subkoviak (1984).

Nótese que en realidad no se está proponiendo un nuevo coeficiente de fiabilidad, sino un método para calcular los ya conocidos coeficientes p_o y *kappa* en el caso de que solo se disponga de una aplicación del test.

El cálculo se va a ilustrar utilizando los datos de la tabla 2.8 en el supuesto de que solo se dispusiera de los datos correspondientes a la forma *A*. Seguiremos los pasos indicados por el propio Subkoviak (1984).

Para proceder al cálculo de p_o y *kappa* se necesitan los datos de la tabla 2.10. A continuación se exponen los pasos para confeccionar la tabla.

1. La primera columna de la tabla 2.10 son las puntuaciones directas (X) obtenidas por las 20 personas en la forma *A* del test (tabla 2.8).
2. La segunda columna son las frecuencias de cada una de las puntuaciones (F_x).
3. La tercera columna (p_x) es una estimación de la proporción de ítems que cabe esperar que responda correctamente una persona que tiene una puntuación determinada X, o, lo que es lo mismo, es su probabilidad de acertar un ítem. Estos valores pueden obtenerse por medio de la fórmula 2.51, aunque otras estimaciones son posibles (Hambleton et al., 1978; Wilcox, 1979). Como señala Subkoviak, cuando el test tiene más de 40 ítems una buena aproximación ya viene dada simplemente por X/n.

$$p_x = \alpha \frac{X}{n} + (1 - \alpha)\left(\frac{\bar{X}}{n}\right) \qquad [2.51]$$

donde:

α: Coeficiente alfa.
X: Puntuación directa.
n: Número de ítems del test.
$\bar{X}$: Media del test.

Veamos su aplicación para los datos de la tabla, teniendo en cuenta que el coeficiente alfa es 0,75 y la media del test 5,3.

$$p_{10} = 0{,}75(10/10) + (1 - 0{,}75)(5{,}3/10) = 0{,}8825$$
$$p_9 = 0{,}75(9/10) + (1 - 0{,}75)(5{,}3/10) = 0{,}8075$$
$$p_8 = 0{,}75(8/10) + (1 - 0{,}75)(5{,}3/10) = 0{,}7325$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

El resto de los valores de la columna se obtienen análogamente.

4. La cuarta columna (P_x) de la tabla 2.10 está formada por los valores obtenidos mediante la distribución binomial. Se procede del siguiente modo. Empezando por la parte superior de la tabla: la probabilidad de que una persona con una puntuación de 10 ($X = 10$), y una probabilidad de acertar cada ítem de 0,88 ($p_x = 0{,}88$), supere siete ítems o más (recuérdese que 7 era el punto de corte establecido) viene dada según las tablas de la distribución binomial por $P_x = 0{,}976$. En el caso siguiente, $X = 9$ y $p_x = 0{,}81$, $P_x = 0{,}896$. Análogamente se obtienen el resto de los valores de la cuarta columna. En suma, se trata simplemente de buscar los valores correspondientes en las tablas de la distribución binomial, teniendo en cuenta el número de ítems (n), el punto de corte establecido (en nuestro ejemplo 7) y la probabilidad de acertar los ítems (p_x).
5. En la columna quinta se reflejan los valores obtenidos mediante la siguiente expresión:

$$P_x^2 + (1 - P_x)^2 \qquad [2.52]$$

Estos valores reflejan la consistencia de la clasificación de cada persona como *master* o *no master,* es decir, superar/no superar el punto de corte. La probabilidad de que una

persona supere el punto de corte en ambas pruebas vendrá dada por P_x^2 y la probabilidad de que falle en ambas será $(1 - P_x)^2$. De modo que la probabilidad de una clasificación consistente fallar-fallar o superar-superar será la suma, que es la expresión 2.52. Veamos la aplicación de esta expresión para los dos primeros casos de la tabla 2.10:

$$(0{,}976)^2 + (1 - 0{,}976)^2 = 0{,}953$$
$$(0{,}896)^2 + (1 - 0{,}896)^2 = 0{,}814$$

Análogamente se obtienen el resto de los valores de la columna.

6. En la sexta columna se estima el número de personas con determinada puntuación que son clasificadas de forma consistente. Los valores de la columna se obtienen multiplicando los de la quinta columna por la frecuencia (F_x). Por ejemplo:

$$(1)(0{,}953) = 0{,}953$$
$$(1)(0{,}814) = 0{,}814$$
$$(3)(0{,}603) = 0{,}1809$$

7. Finalmente, la suma de los valores de la columna séptima es una estimación de las personas que superarán el punto de corte en ambos test. Su cálculo es inmediato, multiplicando los valores de la columna segunda por los de la cuarta (F_x, P_x).

A partir de los datos de la tabla 2.10 ya se pueden calcular los coeficientes p_o y *kappa.*

$$\text{Coeficiente } p_o\text{: } \frac{15{,}232}{20} = 0{,}76$$

Para aplicar la fórmula del coeficiente *kappa* hay que obtener previamente la proporción de coincidencias por azar:

$$P_a = \left(\frac{6{,}655}{20}\right)^2 + \left(1 - \frac{6{,}655}{20}\right)^2 = 0{,}56$$

Aplicando la fórmula del coeficiente *kappa* para proporciones (2.54),

$$K = \frac{P_c - P_a}{1 - P_a} = \frac{0{,}76 - 0{,}56}{1 - 0{,}56} = 0{,}45$$

Método de Huynh

Propuesto por Huynh (1976), es un método elegante matemáticamente para estimar los coeficientes p_o y *kappa.* En su formulación original los cálculos resultan bastante laboriosos, por lo que Peng y Subkoviak (1980) propusieron un método de cálculo

TABLA 2.10

Cálculos requeridos por el método Subkoviak

X	F_x	p_x	P_x	$P_x^2 + (1 - P_x)^2$	$F_x[P_x^2 + (1 - P_x)^2]$	$F_x P_x$
10	1	0,88	0,976	0,953	0,953	0,976
9	1	0,81	0,896	0,814	0,814	0,896
8	3	0,73	0,727	0,603	1,809	2,181
7	2	0,66	0,541	0,503	1,006	1,082
6	2	0,58	0,333	0,556	1,112	0,666
5	3	0,51	0,189	0,693	2,079	0,567
4	2	0,43	0,080	0,853	1,706	0,160
3	4	0,36	0,030	0,942	3,768	0,120
2	1	0,28	0,007	0,986	0,986	0,007
0	1	0,13	0,000	0,999	0,999	0,000
Total	**20**				**15,232**	**6,655**

mucho más sencillo, que es el que seguiremos aquí. Esta variación se basa en el supuesto, por otra parte razonable, de que si se aplicasen dos formas paralelas, la distribución conjunta sería aproximadamente normal. Diversos autores opinan que esta asunción parece plausible cuando el número de ítems es superior a ocho y la media del test dividida entre el número de ítems (X/n) está entre 0,15 y 0,85.

Veamos su aplicación a los datos de la tabla 2.8, suponiendo, como en el caso anterior, que solo se dispone de la forma *A* y que el coeficiente de fiabilidad de esa forma es $KR_{21} = 0{,}70$. Recuérdese que la media del test era 5,3, la varianza 6,41 y que el punto de corte se establecía en 7. A partir de esos datos se procede como sigue.

1. Se calcula la desviación normal correspondiente al punto de corte (C) según la siguiente expresión:

$$Z = \frac{C - 0{,}5 - \bar{X}}{S_x} \qquad [2.53]$$

$$Z = \frac{7 - 0{,}5 - 5{,}3}{\sqrt{6{,}41}} = 0{,}47$$

2. Se busca en las tablas de la curva normal la proporción P_z correspondiente al valor $Z = 0{,}47$. En nuestro caso,

$$P_z = 0{,}68$$

3. Mediante la tabla *F* se obtiene la probabilidad P_{zz} de la distribución conjunta de dos variables normales con correlación $KR_{21} = 0{,}70$, para el valor de $Z = 0{,}47$.

$$P_{zz} = 0{,}58$$

4. Se procede al cálculo de p_o y *kappa:*

$$P_o = 1 + 2(P_{zz} - P_z) \qquad [2.54]$$

$$P_o = 1 + 2(0{,}58 - 0{,}68) = 0{,}80$$

$$k = \frac{P_{zz} - P_z^2}{P_z - P_z^2} \qquad [2.55]$$

$$k = \frac{0{,}58 - 0{,}68^2}{0{,}68 - 0{,}68^2} = 0{,}54$$

Cabe preguntarse cuál de los dos coeficientes (p_o o *kappa*) es preferible utilizar. No existe una respuesta definitiva por parte de los especialistas. Cualquiera de ellos resulta apropiado si se utiliza con prudencia, siendo recomendable proporcionar no solo el valor numérico del coeficiente, sino también otros datos complementarios que puedan resultar útiles para su interpretación precisa, tales como la tabla de frecuencias de la clasificación, la distribución de frecuencias, la media y desviación típica, el punto de corte y errores típicos de medida para distintos niveles de la variable medida.

Como ocurría con el coeficiente de fiabilidad clásico, la longitud del test y la variabilidad de la muestra en el test influyen en ambos coeficientes. Tanto el incremento de la longitud del test como la variabilidad de la muestra tienden a incrementar el tamaño de los coeficientes p_o y *kappa.* Sin olvidar, claro está, que la clave para una buena fiabilidad es que los ítems sean de calidad y constituyan una muestra representativa del dominio a evaluar.

Pero el factor que tiene un mayor efecto sobre el valor de ambos coeficientes es la ubicación del punto de corte que se establezca para llevar a cabo la clasificación de las personas de la muestra. El valor de ambos coeficientes para un mismo test variará en función de dónde se establezca el punto de corte. Por tanto, no debe hablarse sin más del valor de los coeficientes p_o y *kappa* para un test, sino para un test y determinado punto de corte. En general, y asumiendo que la distribución de las puntuaciones del test es unimodal, el valor de p_o tiende a aumentar si el punto de corte se ubica en zonas extremas de la distribución. Con el coeficiente *kappa* ocurre lo contrario: su valor aumenta cuando el punto de corte está cercano a la media de la distribución. Debido a esta clara incidencia del punto de corte, ha recibido bastante atención el estudio de la metodología para su ubicación adecuada, a la que dedicamos un breve apartado más adelante.

Coeficiente de Livingston

Todos los acercamientos al cálculo de la fiabilidad de los TRC vistos hasta ahora tienen en común que parten de una clasificación de las personas en varias categorías y asumen que los errores que se cometen al clasificar son de la misma gravedad, es decir, el error de que personas que dominan la materia

(masters) sean clasificadas eventualmente como *no masters* se considera igual de grave que cuando los *no masters* se clasifican como *masters.* A todos los métodos con esa asunción suele denominárselos como de *pérdida de umbral,* aludiendo a las pérdidas o errores que se producen al clasificar las personas tras establecer un determinado umbral o punto de corte. Sin embargo, los dos índices que se verán a continuación (Livingston, 1972; Brennan y Kane, 1977) no son de ese tipo, pues sí tienen en cuenta la importancia relativa de las clasificaciones incorrectas. Ambos índices van a considerar más graves los errores de clasificación de las personas alejadas del punto de corte que de las ubicadas cerca del mismo. La lógica es clara: es fácil cometer errores cuando las personas están muy cerca del punto de corte, pues cualquier mínimo error puede conducir a equivocarse; pero no hay justificación para equivocarse en la clasificación si la persona está muy alejada del punto de corte. Por tanto, parece razonable no dar el mismo peso a ambos tipos de errores, y eso es lo que hacen los coeficientes de Livingston (1972) y Brennan y Kane (1977), tenerlo en cuenta. Estos coeficientes reciben el nombre genérico de «pérdida de error cuadrático», que alude a que utilizan la distancia cuadrática de las puntuaciones al punto de corte. Cuanto más se alejen las puntuaciones del punto de corte, mayores serán los errores cuadráticos.

El coeficiente de fiabilidad propuesto por Livingston fue desarrollado en el contexto de la teoría clásica de los test, por lo que su terminología resulta familiar, y es muy fácil de calcular. Aunque aquí se presenta dentro del apartado de coeficientes que solo exigen una sola aplicación del test, lo cual es verdad, con una ligera modificación, que luego se verá, puede utilizarse también en el caso de disponer de dos formas paralelas de un test.

La fórmula viene dada por:

$$K^2_{xv} = \frac{\alpha\sigma^2_x + (\mu - c)^2}{\sigma^2_x + (\mu - c)^2} \qquad [2.56]$$

donde:

α: Coeficiente alfa.
σ^2_x: Varianza del test.
μ: Media del test.
c: Punto de corte.

Vamos a aplicar la fórmula a la forma *A* de la tabla 2.8. Recuérdese que el coeficiente alfa era 0,75, la media 5,3, la varianza 6,41 y el punto de corte se establecía en 7.

$$K^2_{xv} = \frac{0{,}75 \times 6{,}41 + (5{,}3 - 7)^2}{6{,}41 + (5{,}3 - 7)^2} = 0{,}83$$

Entre las propiedades de K^2_{xv}, cabe destacar que aumenta con el incremento del coeficiente alfa, pues, como se puede observar en la fórmula, el coeficiente alfa aparece en el numerador. Cuando alfa toma el valor de 1, el coeficiente K^2_{xv} también es 1. Como se desprende directamente de su fórmula, cuando el punto de corte (c) coincide con la media del test (μ), el coeficiente K^2_{xv} es igual a alfa. A medida que el punto de corte se aleja de la media del test, el valor de $(\mu - c)^2$ aumenta, y así lo hace K^2_{xv}, por lo que su valor siempre es igual o mayor que alfa:

$$K^2_{xv} \geqslant \alpha$$

Dado que al ir separándose de la media del test el punto de corte el valor de K^2_{xv} va aumentando, podría argumentarse que, mediante esa lógica, para lograr que un test fuese muy fiable sería cuestión de separar el punto de corte lo suficiente de la media del test. Bien, eso es verdad, igual que era verdad en el enfoque clásico que, si se aumentaba convenientemente la variabilidad de la muestra, se lograba una elevada fiabilidad. Pero ni en uno ni en otro caso tiene sentido operar de ese modo solo por elevar la fiabilidad. Además, el punto de corte se establece allí donde se estima que tiene más sentido y significación, y no donde más conviene para obtener una elevada fiabilidad, igual que ocurría en lo tocante a la variabilidad, que será la que sea en función de la población a la que vaya dirigido el test y no elegida a conveniencia de un estudio de fiabilidad.

Livingston (1972) demuestra de forma elegante que la fórmula de Spearman-Brown clásica que relaciona la longitud y la fiabilidad del test es perfectamente aplicable a su coeficiente en el entorno de los test referidos al criterio:

$$K^2_{xv} = \frac{nK^2_{xv}}{1 + (n - 1)K^2_{xv}} \qquad [2.57]$$

donde n, como en la teoría clásica, es el número de veces que se alarga el test.

Ejemplo

¿Cuál sería el valor del coeficiente de Livingston si se añadiesen cinco ítems a la forma *A* de la tabla 2.8?

$$n = \frac{10 + 5}{10} = 1{,}5$$

$$K^2_{xv} = \frac{1{,}5 \times 0{,}83}{1 + (1{,}5 - 1)0{,}83} = 0{,}88$$

El valor obtenido con 10 ítems (0,83) pasa a 0,88 cuando se añaden cinco ítems. En el caso de que se disponga de dos formas paralelas o de dos aplicaciones del mismo test, la fórmula propuesta en 2.56 viene dada por:

$$K_{xx'} = \frac{\rho_{xx'}\sigma_x\sigma_{x'} + (\mu_x - c)(\mu_{x'} - c)}{\sqrt{[\sigma^2_x + (\mu_x - c)^2][\sigma^2_{x'} + (\mu_{x'} - c)^2]}} \qquad [2.58]$$

donde X, X' son las dos formas paralelas del test, y $\rho_{xx'}$, la correlación entre ellas, es decir, el coeficiente de fiabilidad. A modo de ejercicio, trate el lector de aplicar esta fórmula a los datos de la tabla 2.8.

Coeficiente de Brennan y Kane

Utilizando datos obtenidos al aplicar el modelo de la teoría de la generalizabilidad, Brennan y Kane (1977) han propuesto un coeficiente muy similar al de Livingston, si bien da resultados algo más bajos que aquel cuando se aplica a los mismos datos. Su fórmula viene dada por:

$$M(C) = \frac{\sigma^2_p + (\mu_{pi} - C)^2}{\sigma^2_p + (\mu_{pi} - C)^2 + \dfrac{\sigma^2_i + \sigma^2_e}{n_i}} \qquad [2.59]$$

Los distintos términos de la fórmula han sido descritos en el epígrafe 8 dedicado a la teoría de la generalizabilidad. El único nuevo es C, que se refiere al punto de corte, expresado en términos de proporción de ítems que han de ser respondidos correctamente para superar la prueba. Nótese que también μ_{pi} ha de expresarse en forma de proporción, siendo la proporción media total de ítems superados por las personas. Véase en el apartado siguiente la aplicación de la fórmula 2.59 a los datos de la tabla 2.11.

Estimación del dominio

Todos los indicadores de fiabilidad de los test referidos al criterio vistos hasta ahora, de una manera u otra, trataban de comprobar en qué medida las clasificaciones hechas a partir del establecimiento de un punto de corte resultaban consistentes; de ahí que muchos autores prefieran denominarlos índices de acuerdo, en vez de coeficientes de fiabilidad propiamente dichos. No debe extrañar que la mayoría de los coeficientes vayan dirigidos a ese fin, pues en la práctica educativa y profesional suele exigirse el establecimiento de un punto de corte entre las personas que superan y no superan el dominio. No obstante, cabe plantearse el problema de forma más general, y preguntarse en qué medida las puntuaciones en la prueba representan las puntuaciones del dominio. O, referido a una persona concreta, preguntarse cuál será su puntuación en el dominio, conocida su puntuación en la prueba, tal como se hacía en la teoría clásica. En el marco de la teoría de la generalizabilidad puede hallarse una respuesta apropiada a esas cuestiones, si bien también otros acercamientos son posibles (Berk, 1984b; Brennan, 1980; Lord y Novick, 1968). Una primera posibilidad, que ya resultará familiar al lector, sería utilizar el error típico de medida clásico para establecer un intervalo confidencial en torno a la puntuación empírica que permita estimar el valor verdadero en el dominio, al modo en que se hacía para los test referidos a normas. También se pueden utilizar para este menester errores típicos de medida basados en la distribución binomial.

Dentro del marco de la teoría de la generalizabilidad, si lo que interesa es llevar a cabo estimaciones de lo que una persona obtendría en el dominio, puede utilizarse la varianza error correspondiente al

caso de decisiones absolutas, que como se vio en el epígrafe 8 viene dado por:

$$\sigma_{e(g)} = \sqrt{\frac{\sigma_i^2 + \sigma_e^2}{n}} \qquad [2.60]$$

A partir de esa varianza error, un coeficiente de generalizabilidad apropiado para el caso de los test referidos al criterio vendría dado por:

$$\rho_{gD}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_i^2 + \sigma_e^2}{n}} \qquad [2.61]$$

donde n es el número de ítems, y los valores de σ_p^2, σ_i^2 y σ_e^2, como se vio en el epígrafe 8, pueden calcularse en función de las medias cuadráticas según las fórmulas:

$$\sigma_p^2 = \frac{MC_p - MC_r}{n_i}$$

$$\sigma_i^2 = \frac{MC_i - MC_r}{n_p}$$

$$\sigma_e^2 = MC_r$$

También se puede utilizar el coeficiente de generalizabilidad visto para los test referidos a normas y que venía dado por:

$$\rho_G^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{n}} \qquad [2.62]$$

Como ya habrá advertido el lector que no se haya saltado la lectura del apartado dedicado a la TG, las fórmulas precedentes provienen de un diseño de una sola faceta (los ítems) cruzada con las personas ($p \times i$), pues todas las personas han de responder a todos los ítems. Ya decíamos allí que en la práctica este es el diseño más habitual.

Ejemplo

En la tabla 2.11 aparecen los resultados de aplicar un test de cinco ítems a 10 personas.

A partir de la tabla 2.11 se elabora mediante análisis de varianza la tabla 2.12. Para ver cómo se hace paso a paso, puede consultarse el ejemplo desarrollado con detalle al tratar los diseños de una sola faceta (subepígrafe 8.3).

TABLA 2.11

Personas	Ítems					$\bar{X}_p$
	1	2	3	4	5	
a	0	1	1	0	0	0,4
b	1	1	0	0	0	0,4
c	0	0	0	0	0	0,0
d	1	1	1	0	0	0,6
e	1	0	0	1	1	0,6
f	1	1	0	0	0	0,4
g	1	1	1	1	0	0,8
h	1	0	1	0	0	0,4
i	1	1	1	1	1	1,0
j	1	0	1	0	0	0,4
$\bar{X}_i$	0,8	0,6	0,6	0,3	0,2	0,5

TABLA 2.12

Fuentes de variación	Suma de cuadrados	Grados de libertad	Medias cuadráticas	Varianza estimada
Personas (p)	3,30	9	0,37	0,036
Ítems (i)	2,40	4	0,60	0,041
Residual (r)	6,80	36	0,19	0,190

A partir de los datos de la tabla 2.12 se calculan los valores del error típico de medida y de los coeficientes expuestos. Empecemos por el error típico de medida:

$$\sigma_{e(g)} = \sqrt{\frac{\sigma_i^2 + \sigma_e^2}{n}} = \sqrt{\frac{0{,}041 + 0{,}190}{5}} = 0{,}21$$

Mediante este valor, y adoptando un determinado nivel de confianza, pueden establecerse intervalos confidenciales en torno a las puntuaciones de las personas, y así estimar la que corresponde en el dominio.

El coeficiente de generalizabilidad para los TRC vendría dado por:

$$\rho_{gD}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_i^2 + \sigma_e^2}{n}} = \frac{0{,}036}{0{,}036 + \frac{0{,}041 + 0{,}190}{5}} = 0{,}44$$

El coeficiente de generalizabilidad correspondiente a los test normativos:

$$\rho_G^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{n}} = \frac{0{,}036}{0{,}036 + \frac{0{,}190}{5}} = 0{,}49$$

El coefiente de Brennan y Kane vendría dado por:

$$M(C) = \frac{0{,}036 + (0{,}5 - 0{,}7)^2}{0{,}036 + (0{,}5 - 0{,}7)^2 + \frac{0{,}041 + 0{,}19}{5}} = 0{,}62$$

9.3. Establecimiento del punto de corte

Como se ha señalado en el apartado anterior al exponer los métodos para estimar la fiabilidad de los TRC, la mayoría de las situaciones aplicadas conllevan el establecimiento de un punto de corte que permita clasificar a las personas en dos grupos (a veces más): aquellas que dominan el criterio evaluado y las que no lo dominan. En los distintos métodos expuestos para calcular la fiabilidad se asumía que el punto de corte ya estaba establecido, pero ¿cómo se establece? Por ejemplo, si se trata de un TRC cuyo objetivo es evaluar los conocimientos de inglés para acceder a una beca en el extranjero, ¿cómo establecer el punto a partir del cual se considera que los estudiantes dominan suficientemente el inglés?, ¿cuánto es suficiente? La respuesta es que dicho punto de corte deben establecerlo jueces expertos en la materia. Ahora bien, superado el problema de contar con los jueces adecuados en calidad y número, existen distintos procedimientos que estos pueden seguir para decidir el punto de corte más apropiado. Ese tipo de procedimientos son los que se expondrán en este apartado. No existe un punto de corte mágico y correcto a priori, pues depende de los jueces, pero, como se irá viendo, ello tampoco quiere decir que cualquier procedimiento seguido sea igualmente válido. El asunto puede parecer un poco obvio, pero cualquiera que haya tratado con un grupo de jueces expertos a los que se encomienda fijar un punto de corte habrá comprobado la dificultad práctica de la tarea. El establecimiento de los puntos de corte de forma adecuada es de suma importancia, pues a menudo determina el futuro profesional de las personas en todo tipo de exámenes y certificaciones. Existe además una tendencia creciente a que la mayoría de las profesiones tengan que certificarse y pasar cada cierto tiempo pruebas para demostrar que siguen al día en sus respectivos campos, lo cual suele hacerse mediante TRC, que obligan a establecer los puntos de corte correspondientes para decidir quién es competente y quién no lo es.

Suele hablarse de puntos de corte relativos y absolutos. Se denominan relativos cuando se fijan en función del grupo de personas evaluadas, y absolu-

tos, cuando solo dependen de la materia evaluada. Los puntos de corte relativos no tienen mucho sentido en el contexto de los TRC, puesto que el objetivo de estos es determinar el dominio que las personas tienen del criterio y no su posición respecto del resto de los componentes del grupo. Si se adopta un procedimiento relativo, se está predeterminando a priori que algunas personas no serán clasificadas como *masters,* cuando es perfectamente posible que todas ellas dominen el criterio. Por ejemplo, sería inadmisible que un profesor estableciese el punto de corte para aprobar a los alumnos en la nota media de los presentados, pues estaría estableciendo de antemano que van a suspender en torno a un 50% de los estudiantes. El aprobado ha de establecerse en función de la materia, no del grupo. Algo bien distinto ocurre en situaciones de selección de personal, oposiciones y, en general, cuando existen muchos más aspirantes que plazas libres, en cuyo caso sencillamente se admite a los mejores, supuesto, claro está, que superen los mínimos exigidos.

Por todo lo dicho anteriormente, aquí nos centraremos fundamentalmente en los puntos de corte de carácter absoluto. No obstante, en el último apartado se tratarán dos métodos que intentan llegar a un compromiso entre este enfoque absoluto y algunos datos de carácter relativo. Se exponen los procedimientos más habituales y clásicos, pero el lector interesado en ir algo más allá puede consultar la abundante bibliografía existente, recomendándose en especial los trabajos de Livingston y Zieky (1982), Berk (1986), Jaeger (1989), Cizek (1996, 2012), Hambleton y Pitoniak (2006) o Zieky et al. (2008), entre otros muchos. Se describen tres bloques de procedimientos: los centrados en el test, los centrados en las personas y los de compromiso.

Procedimientos centrados en el test

Los procedimientos descritos en este apartado para fijar el punto de corte se basan en los juicios de los expertos acerca de los distintos ítems del test, lo que explica su denominación de centrados en el test. Se describirán los métodos propuestos por Nedelsky (1954), Angoff (1971) y Ebel (1972). Los tres métodos requieren seleccionar los jueces apropiados en calidad y número, así como darles un cierto entrenamiento y formación. No se entra aquí en cómo llevar a cabo esas tareas; únicamente señalar que la muestra de jueces debe ser representativa y cuanto más amplia mejor. Temas interesantes debatidos por los expertos son si es preferible que los jueces trabajen en grupo o individualmente, si se les debe obligar a llegar a consensos, si hay que proporcionarles la alternativa correcta o no, etc. El análisis de esta problemática excede los propósitos de nuestra sucinta exposición, pero el lector interesado encontrará abundante material en las referencias bibliográficas citadas. Los tres métodos descritos a continuación varían en el tipo de tareas solicitadas de los jueces y en cómo se procesan y organizan los juicios emitidos por estos.

Método de Nedelsky

Propuesto por Nedelsky (1954), solo se puede utilizar cuando los ítems son de elección múltiple, pues requiere que los jueces analicen cada una de las alternativas. Una vez analizadas, deben decidir cuáles consideran que serían detectadas como erróneas por una persona que tuviese los conocimientos mínimos exigibles para dominar el criterio. Por ejemplo, considérese el siguiente ítem de geografía.

La capital de Estados Unidos es: Nueva York, Washington, Montreal, San Francisco, Ottawa. Un juez podría considerar que una persona con los mínimos conocimientos geográficos exigibles para aprobar sería capaz de descartar como erróneas Montreal, San Francisco y Ottawa, pero no Nueva York. En función de las respuestas de todos los jueces a todos los ítems, se establece el punto de corte, o conocimientos mínimos exigibles para aprobar o superar la prueba. Veamos en concreto cómo se tienen en cuenta las opiniones de los jueces para establecer el punto de corte.

En el método de Nedelsky se asume que ante un ítem las personas primero descartan las alternativas que consideran claramente erróneas y luego eligen al azar entre las restantes. Bajo esta óptica, la puntuación esperada de una persona en un ítem viene dada dividiendo la unidad por el número de alternativas no descartadas. En el ejemplo anterior, el valor esperado sería 1/2 = 0,5, pues quedaban sin descartar dos alternativas, Nueva York y Washington. Para obtener el valor esperado para todo el test se suman los valores esperados de cada ítem. De modo que tendremos un valor esperado del test para cada juez. ¿Cómo combinar los valores de los distintos jueces para obtener el punto de corte único? Lo más habitual es calcular

la media o mediana de los valores asignados por los distintos jueces. También se pueden eliminar los valores extremos antes de calcular la media o la mediana. Si la variabilidad entre los jueces es alta, lo cual no es deseable, pues implica poca fiabilidad interjueces, la mediana es más indicada.

Cuando se va a utilizar la fórmula para corregir los efectos del azar (véase el epígrafe 5 del capítulo 4), también debe corregirse con ella el punto de corte. Por ejemplo, si en un test de 10 ítems de cinco alternativas cada uno la media de los valores esperados de cuatro jueces resultó ser 6, el punto de corte sin corregir los efectos del azar se establece en 6. Pero si se corrigen estos mediante la fórmula expuesta en el epígrafe 5 del capítulo 4, el punto de corte sería:

$$6 - \frac{(10 - 6)}{5 - 1} = 5$$

En general, el método de Nedelsky funciona bien, aunque no carece de limitaciones. Por ejemplo, el valor esperado de un ítem nunca puede tomar valores entre 0,50 y 1, pues o bien solo quedan dos alternativas sin descartar, en cuyo caso el valor esperado es 0,5, o queda solo una, en cuyo caso es 1. Como señala Shepard (1980), este método tiende a dar valores más bajos para el punto de corte que el resto de los métodos, debido a la resistencia habitual de los jueces a considerar que todas las personas responderían correctamente el ítem.

Método de Angoff

Propuesto por Angoff (1971), es muy parecido al de Nedelsky visto en el apartado anterior, si bien tiene la gran ventaja de que permite su aplicación a todo tipo de ítems, aunque no sean de elección múltiple. En este método no se pide a los jueces que analicen cada una de las alternativas de los ítems, como se hacía en el de Nedelsky; aquí los jueces emiten evaluaciones globales de cada ítem. Se les pide que digan cuál es la probabilidad de que una persona con los conocimientos mínimos exigibles supere el ítem. A veces resulta más fácil a los jueces si se les plantea esta cuestión de forma ligeramente distinta, preguntándoles cuántas de entre 100 personas con los conocimientos mínimos exigibles superarían el ítem. Una vez asignadas las probabilidades a cada ítem, la suma de estas da el punto de corte exigible para superar el criterio. Si hay varios jueces, como es habitual, para obtener el punto de corte único se combinan sus puntuaciones calculando la media o la mediana, igual que en el caso del método de Nedelsky. También se procede como allí en el caso de corrección de los efectos del azar. Nótese que si los ítems son de elección múltiple, las probabilidades asignadas por los jueces deberían ser al menos iguales o superiores a la correspondiente por mero azar, es decir, la unidad dividida entre el número de alternativas del ítem.

EJEMPLO

En la tabla adjunta aparecen las probabilidades de que las personas con una competencia mínima exigible superen los ítems de un test. Veamos cómo se establece el punto de corte por el método de Angoff sin corregir los efectos de azar y corrigiéndolos.

Ítems	Juez A	Juez B	Juez C
1	0,50	0,50	0,50
2	0,33	0,50	0,33
3	0,25	0,25	0,25
4	1,00	0,50	0,50
5	0,25	0,33	0,25
Total	**2,33**	**2,08**	**1,83**

La media de los tres jueces es 2,08, por lo que si la puntuación final del test se obtiene sin utilizar la fórmula de corrección de los efectos del azar, el punto de corte estaría ubicado en ese valor de 2,08.

Si para obtener las puntuaciones de las personas en test se corrigen los efectos del azar, el punto de corte vendría dado por:

$$2{,}08 - \frac{(5 - 2{,}08)}{5 - 1} = 1{,}35$$

(Para una exposición de la lógica y fórmula de la corrección de los efectos del azar, véase el epígrafe 5 del capítulo 4.)

El método de Angoff es el más utilizado en la práctica, pues resulta sencillo de explicar a los jueces y fácil de utilizar en la mayoría de las situaciones. Aquí no nos ocupamos de cómo proceder con los

jueces para llegar a la asignación de las probabilidades finales, es decir, cómo deben trabajar, individualmente, en grupo, por consenso, instrucciones, número de jueces, etc. La utilización de un procedimiento u otro puede variar en función de cada situación específica, pero en general cualquiera de las opciones es válida si se procede con rigor metodológico.

A veces se utiliza una variante del método de Angoff, consistente en proporcionar a los jueces una serie de probabilidades entre las que tienen que elegir, en vez de generarlas ellos mismos. Livingston y Zieky (1982) desaconsejan esta variante, pues limita las opciones de los jueces, y además consideran que va en contra del espíritu del método de Angoff.

Jaeger (1982) también propuso otra variante del método, consistente en preguntar a los jueces si el ítem debería ser contestado correctamente por todos los sujetos. En un proceso iterativo se permite a los jueces ir modificando sus juicios en función de la información que se les facilita, tal como sus juicios previos, de las opiniones de otros jueces y de la dificultad de los ítems. El punto de corte es la mediana más baja de los diferentes grupos de jueces finalizadas las iteraciones. Un problema de esta variante es que solo permite a los jueces asignar probabilidades de 0 o 1 a los ítems, pues o consideran que lo superan o que lo fallan.

Método de Ebel

Propuesto por Ebel (1972), puede considerarse un método que requiere dos pasos: primero los jueces clasifican los ítems según su dificultad y luego según su relevancia. Ebel sugiere que se utilicen tres niveles de dificultad (fácil, medio y difícil) y cuatro niveles de relevancia (fundamental, importante, aceptable y dudoso). De este modo los ítems aparecen clasificados en una tabla de 3 × 4, que da lugar a 12 categorías distintas. Una vez clasificados los ítems en las 12 categorías, se pide a los jueces que para cada una de ellas señalen el porcentaje de personas que teniendo una competencia mínima exigible superarían los ítems similares a los de la categoría contemplada. La suma ponderada en función del número de ítems de cada categoría es el punto de corte del test. Las puntuaciones de los distintos jueces se combinan como en los métodos anteriores, calculando la media o la mediana.

Ejemplo

En la tabla adjunta aparecen los 25 ítems de un test clasificados por un juez en las 12 categorías sugeridas por Ebel, así como los juicios emitidos para los ítems de cada una de las 12 categorías.

	Fáciles	Medios	Difíciles
Fundamentales	Ítems: 1, 3 Juicio: 90% superan	Ítems: 2, 5 Juicio: 80% superan	Ítems: 4 Juicio: 60% superan
Importantes	Ítems: 7, 10, 8 Juicio: 80% superan	Ítems: 6, 9, 25 Juicio: 70% superan	Ítems: 11, 24 Juicio: 40% superan
Aceptables	Ítems: 14, 19, 23 Juicio: 95% superan	Ítems: 16, 18, 22 Juicio: 90% superan	Ítems: 0
Dudosos	Ítems: 13, 20 Juicio: 90% superan	Ítems: 15, 17, 21 Juicio: 50% superan	Ítems: 12 Juicio: 30% superan

Para obtener el punto de corte se multiplican las proporciones de cada casilla por el número de ítems de la casilla y se suma:

$$2(0{,}90) + 2(0{,}80) + 1(0{,}60) + 3(0{,}80) + 3(0{,}70) + \\ + 2(0{,}40) + 3(0{,}95) + 3(0{,}90) + 0 + 2(0{,}90) + \\ + 3(0{,}50) + 1(0{,}30) = 18{,}45$$

Si existiesen más jueces, se combinarían sus puntuaciones como se señaló en los métodos anteriores, utilizando la media o la mediana, bien eliminando las puntuaciones extremas o sin eliminarlas.

Procedimientos centrados en las personas

Los procedimientos expuestos en el apartado anterior se basaban en los juicios de los expertos sobre los ítems del test. Sin embargo, los métodos descritos en este apartado se valen de las opiniones

de los jueces sobre la competencia de las personas. Para aplicar estos métodos hay que disponer de jueces expertos en la materia a evaluar, y que además conozcan perfectamente la competencia de las personas en la materia objeto de evaluación. Recuérdese que en los procedimientos centrados en el test los jueces solo tenían que ser expertos en la materia a evaluar, pero no necesitaban conocer la competencia de las personas. A continuación se describen los dos métodos más clásicos: el del grupo límite (Zieky y Livingston, 1977) y el de los grupos de contraste (Berk, 1976).

Método del grupo límite

En este método, propuesto por Zieky y Livingston (1977), se pide a los jueces que identifiquen de entre las personas a las que va destinado el test aquellas que en su opinión estarían en el límite de superarlo, es decir, aquellas cuyos conocimientos en la variable medida no son del todo inadecuados, pero tampoco completamente adecuados. En suma, se les pide que identifiquen las personas que están en el límite de lo exigible, situadas entre las que dominan el criterio y las que no lo dominan. Una vez identificado por los jueces un grupo de personas de estas características, se les aplica el test en el cual se está interesado en fijar el punto de corte. El valor de este será la media o la mediana de las puntuaciones de este grupo de personas en el test. Se utiliza con más frecuencia la mediana, ya que es menos sensible a la variabilidad de las puntuaciones. Si el test se aplica antes de que los jueces emitan sus valoraciones, no se les deben dar a conocer las puntuaciones de las personas, pues podrían influir en sus juicios.

El método resulta muy sencillo de aplicar en la práctica, y funciona bien, siempre y cuando, claro está, la detección del grupo de casos límites por los jueces sea correcta. Nótese que solo se puede aplicar en situaciones en las que se dispone de personas conocidas por los jueces en la variable de interés, lo cual suele ocurrir con frecuencia en ambientes académicos, donde los profesores conocen a los alumnos, o en el ámbito profesional, donde los supervisores conocen la competencia profesional de los trabajadores. Un inconveniente con el que se tropieza con frecuencia al aplicarlo es que el grupo de personas límite no sea lo suficientemente amplio.

Ejemplo

Una empresa juguetera ha dado un curso de formación técnica a 50 trabajadores en paro para enseñarles a construir un nuevo juguete que piensa lanzar al mercado la próxima campaña. Después del curso de formación, en opinión de los monitores 11 de las personas que asistieron al curso han conseguido una formación que consideran límite. Sometidas las 50 personas asistentes a una prueba sobre su competencia para elaborar el nuevo juguete, las puntuaciones de las 11 del grupo considerado límite por los jueces fueron las siguientes: 20, 21, 22, 23, 23, 24, 25, 25, 26, 26, 60. ¿Dónde se establecerá el punto de corte?

Dado que la mediana de las 11 puntuaciones es 24, el punto de corte se establecería en 24. En este caso la media no sería apropiada, pues hay un caso extremo (60) que arrastra la media hacia arriba. Si se suprimen las dos puntuaciones extremas (20 y 60), la media sería 23,9, que coincide prácticamente con la mediana.

Método de los grupos de contraste

Su utilización en este ámbito fue sugerida por Berk (1976), si bien la lógica del método era bien conocida y utilizada en el estudio de la validez de las pruebas. Como ocurría con el método anterior del grupo límite, para aplicar este método hay que disponer de jueces que conozcan bien el rendimiento de las personas en el dominio de interés que se pretende medir con la prueba en la cual se está interesado en establecer el punto de corte. Una vez que se dispone de los jueces adecuados, se les pide que clasifiquen a las personas a las que se aplicará el test en dos grupos: las que en su opinión dominan el criterio y las que no lo dominan. Por lo general, no se puede trabajar con toda la población, hay que utilizar una muestra representativa que sea lo más amplia posible. Clasificadas de ese modo las personas por los jueces y conocidas sus puntuaciones en el test, se trata de elegir el punto de corte que separe con la mayor precisión posible a las que dominan el criterio de las que no lo dominan. El caso ideal sería aquel en el que todos los clasificados por los jueces como *masters* en el dominio estuviesen por encima del punto de corte elegido, y todos los clasificados como *no masters,* por debajo. En la

práctica ese caso ideal no se suele dar, por lo que hay que elegir un punto de corte que minimice los errores de clasificación.

La gama de posibilidades para determinar el punto de corte es amplia. Un método gráfico muy sencillo consiste en representar gráficamente ambas distribuciones, la del grupo de aquellos que según los jueces superarían la prueba y los que no, y elegir como punto de corte la intersección de ambas distribuciones, según se observa en la figura 2.1.

Si en la figura 2.1 se mueve el punto de corte hacia la derecha, se reducen los falsos positivos, es decir, se reduce la probabilidad de considerar *masters* a los que no lo son. Por el contrario, si el punto de corte se mueve hacia la izquierda, se reducen los falsos negativos, es decir, la probabilidad de considerar *no masters* a quienes realmente lo son. Es importante tener esto en cuenta, pues pueden existir situaciones prácticas en las que interese más minimizar un tipo de error que otro.

Otro método sencillo y muy utilizado es el sugerido por Livingston y Zieky (1982). Se divide a las personas en varias categorías según su puntuación en el test y se computa para cada categoría el número de personas que los jueces consideraron *masters* y *no masters*. Siguiendo la misma lógica que los métodos psicofísicos clásicos para determinar el umbral absoluto, se elige como punto de corte aquella puntuación que deja por debajo el 50% de los casos considerados *masters* por los jueces.

Ejemplo

Un grupo de 103 estudiantes fueron clasificados por sus profesores en aprobados y suspensos. Tras aplicar a los 103 estudiantes un test referido al criterio de 25 ítems, la distribución de sus puntuaciones y las opiniones de los profesores aparecen en la tabla adjunta.

Test	*F*	Aprobados	Suspensos	Porcentaje aprobados
21-25	10	10	0	100
16-20	20	15	5	80
11-15	30	14	16	50
6-10	34	10	24	22
0-5	9	1	8	2
Total	**103**	**50**	**53**	

En la última columna se observa que la puntuación que deja por debajo al 50% de los clasificados por los jueces como aprobados es 15; luego ese será el punto de corte elegido. En la mayoría de los ca-

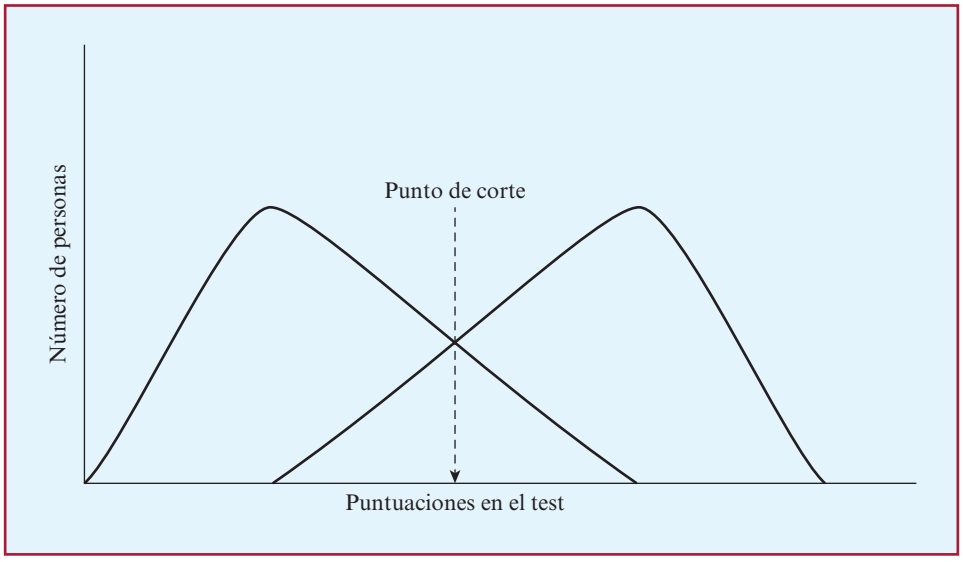


Figura 2.1.—Método de los grupos de contraste.

sos la puntuación que deja por debajo al 50% no aparecerá directamente como en este ejemplo, por lo que habrá que interpolar, de la misma forma que se hace para calcular los percentiles. También se puede utilizar alguna de las estrategias habituales en psicofísica clásica para determinar el umbral absoluto; véanse, por ejemplo, Blanco (1996) o Muñiz (1991). Autores como Livingston y Zieky (1982) proponen que se suavicen los porcentajes acumulados para aumentar su estabilidad, y sugieren métodos sencillos de carácter gráfico, o medias móviles, para llevarlo a cabo. No obstante, si la muestra es suficientemente grande y los intervalos no se hacen demasiado estrechos, puede procederse directamente, como se hizo en el ejemplo. Esta forma de ubicar el punto de corte asume que todos los tipos de errores cometidos al clasificar tienen la misma importancia, pero otras muchas opciones han sido propuestas por diversos autores (Koffler, 1980; Peters, 1981).

Un inconveniente de este método de grupos de contraste es que exige mucho tiempo, pues los jueces tienen que juzgar una a una a todas las personas de la muestra. Para evitar este inconveniente, Livingston y Zieky (1982) apuntan la posibilidad de utilizar el método arriba-abajo, también usado en psicofísica, consistente en ir presentando a los jueces únicamente personas cercanas (por encima y por debajo) al previsible punto de corte, con lo cual se ahorra tiempo, si bien la aplicación correcta conlleva serias dificultades. Remitimos a los textos de psicofísica citados para una exposición del método.

Procedimientos de compromiso

Los procedimientos expuestos hasta ahora para fijar el punto de corte se consideran de carácter absoluto, pues en todos ellos los jueces establecen un mínimo de conocimientos que una persona necesita para superar el criterio, independientemente de lo que haga el resto de las personas del grupo. El que una persona superase o no el criterio no estaba en función de su posición relativa en el grupo, de modo que, fijado el punto de corte, podía darse el caso extremo de que todas o ninguna de las personas evaluadas lo superasen.

Los métodos que se exponen en este apartado utilizan tanto la información de carácter absoluto como la relativa al grupo, tratando de llegar a un compromiso combinando ambos tipos de datos. Esta lógica seguramente extrañará al lector, pues al introducir el problema de los puntos de corte ya se indicó que el dominio o no de un criterio por parte de una persona solo debería depender de sus conocimientos en relación con el punto de corte establecido, no de lo que hiciesen los demás. Desde un punto de vista teórico, esto sigue siendo correcto, pero en muchas situaciones aplicadas el establecimiento de puntos de corte o estándares tiene unas repercusiones sociales tan fuertes que excede los planteamientos puramente psicométricos. Piénsese, por ejemplo, en las implicaciones que tendría en España el establecimiento de unos estándares profesionales que tuviesen que superar cada cierto número de años los funcionarios para continuar en su puesto. En estos procesos tan complejos no solo estarían interesados los expertos en la medición, habría que tener en cuenta a los ciudadanos que pagan la Administración con sus impuestos, a los poderes públicos, a los propios funcionarios implicados, etc. En España esto aparece como lejano, pero en otros países de nuestro entorno occidental se están llevando a cabo rutinariamente procesos de certificación de este tipo para distintas profesiones. En estas circunstancias, conviene tener en cuenta no solo los criterios absolutos, sino también su relación con las poblaciones reales a evaluar. Si solo se atendiesen los criterios absolutos, podría ocurrir que en ocasiones resultasen poco realistas, fallando en alcanzarlos la mayoría del colectivo objeto de la evaluación o, por el contrario, superándolos todo el mundo. Esto, que en teoría no representa ningún problema, plantea situaciones engorrosas y de credibilidad en la práctica; de ahí estos métodos que tratan de llegar a un compromiso, combinando los datos de carácter absoluto y relativo.

Se exponen dos de los métodos que más atención han recibido (Beuk, 1984; Hofstee, 1983), aunque existen otros (De Gruijter, 1980; Grosse y Wright, 1986). Para una buena revisión de estos métodos, véanse De Gruijter (1985) o Milis y Melican (1988).

Método de Beuk

Fue propuesto por Beuk en 1984 y en él los jueces tienen que responder a dos preguntas cuyas respuestas se utilizarán luego para establecer el punto de corte de compromiso. La primera de las pregun-

tas aporta datos de carácter absoluto, y la segunda, relativos. Teniendo en cuenta las respuestas de todos los jueces a ambas preguntas y los resultados empíricos de las personas en el test, se establecerá un punto de corte de compromiso entre los tres tipos de información. Las preguntas que se formulan a los jueces son las siguientes:

1. Porcentaje mínimo de ítems de la prueba que deben responder correctamente las personas para superarla.
2. Porcentaje de personas que superarán la prueba.

Como se puede ver, la primera de las cuestiones alude, como en los métodos previos, a datos absolutos, y la segunda es de carácter relativo.

Una vez aplicada la prueba a las personas, estas estimaciones de los jueces se contrastan con los datos empíricos obtenidos y se llega a un compromiso entre las tres fuentes de datos para establecer el punto de corte. Ahora bien, existen muchas formas posibles de compromisos. La de Beuk, que se expone a continuación, es muy razonable, pero otras muchas son pensables. Para aplicar el método de Beuk se procede como se ilustra en la figura 2.2.

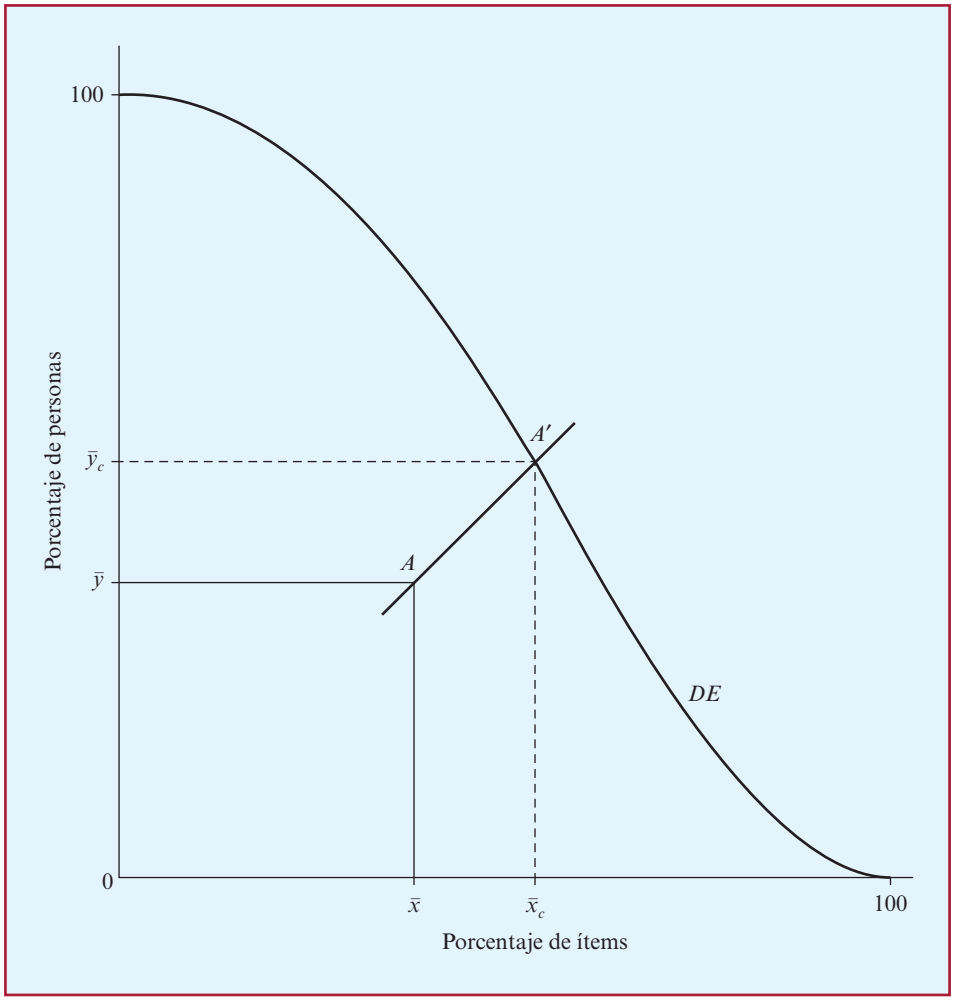


Figura 2.2.—Método de Beuk.

1. Se trazan dos ejes, colocando en abscisas los porcentajes de ítems que en opinión de los jueces hay que contestar correctamente para superar la prueba, es decir, las respuestas de los jueces a la primera pregunta que se les hace.
2. Se calculan las medias de las respuestas de los jueces a las dos preguntas ($\overline{X}$, $\overline{Y}$) y se obtiene el punto *A*.
3. Se obtiene la distribución empírica (*DE*) de las puntuaciones de las personas en el test. Lógicamente resulta decreciente, pues a medida que se va exigiendo superar más ítems para pasar el test (eje de abscisas), decrece el porcentaje de personas que lo pasan (eje de ordenadas).
4. Se obtiene el punto *A'*, intersección de la recta *AA'* con la distribución empírica (*DE*). Para ello se hace pasar por *A* una recta cuya pendiente es el cociente entre la desviación típica de las respuestas de los jueces a la pregunta 2 y la desviación típica de sus respuestas a la primera pregunta, es decir, S_y/S_x. La fórmula de dicha recta viene dada por: $Y = (S_y/S_x)(X - \overline{X}) + \overline{Y}$. Las razones expuestas por Beuk para asignar esta pendiente a la recta resultan coherentes con la idea de compromiso del método, y según los datos que aporta funciona bien. Recuérdese que, en general, para dos variables *X* e *Y,* la pendiente de la recta de regresión de *Y* sobre *X* según el criterio de mínimos cuadráticos viene dada por $r_{xy}(S_y/S_x)$.
5. Para obtener el punto de corte de compromiso se proyecta *A'* sobre el eje de abscisas, obteniéndose el punto de corte (X_c), expresado en forma de porcentaje de ítems del test que se han de superar. Si se prefiere expresar en términos del número de ítems del test, se multiplica este valor (X_c) por el número de ítems (*n*) que tenga el test: (*Xc*)(*n*).

Método de Hofstee

El método de Hofstee (1983) constituye un compromiso entre la información proporcionada por los jueces a cuatro preguntas y la distribución empírica de las puntuaciones de las personas en el test. Las cuatro preguntas son las siguientes:

1. Punto de corte que consideran satisfactorio, aunque lo superen todas las personas ($PC_{máx}$). Se establece en términos del porcentaje de ítems que han de superarse.
2. Punto de corte insatisfactorio, aunque no lo supere nadie ($PC_{mín}$).
3. Porcentaje máximo admisible de personas que fallan en la prueba ($F_{máx}$).
4. Porcentaje mínimo admisible de personas que fallan en la prueba ($F_{mín}$).

Con la información obtenida en esas cuatro preguntas y la distribución empírica (*DE*) de los resultados en el test, se procede como se ilustra en la figura 2.3. En el eje de abscisas aparecen los porcentajes de ítems respondidos correctamente; en el de ordenadas, los porcentajes de personas que no superan los ítems correctos exigidos. La distribución empírica (*DE*) refleja cómo, a medida que se exige un mayor porcentaje de ítems correctos para superar el test (abscisas), aumenta el porcentaje de personas que fallan, es decir, que no superan la prueba (ordenadas).

1. Se obtienen los puntos *A* y *B,* como se indica en el gráfico: $A(PC_{mín}, F_{máx})$ y $B(PC_{máx}, F_{mín})$.
2. Se unen mediante una recta los puntos *A* y *B*.
3. El punto de intersección de la recta *AB* y la distribución empírica (*DE*) se proyecta sobre el eje de abscisas, obteniéndose así el punto de corte de compromiso (PC_c). Si en vez de porcentajes se prefiere utilizar el número de ítems, se multiplica el valor de PC_c por el número de ítems (*n*) del test:

$$(PC_c)(n)$$

Mediante el método de Hofstee se llega a una solución de compromiso entre la información absoluta y la relativa. Aunque no es fácil que ocurra en la práctica, cabe la posibilidad de que la recta *AB* no se encuentre con la distribución empírica (*DE*), en cuyo caso el método no proporcionaría una solución.

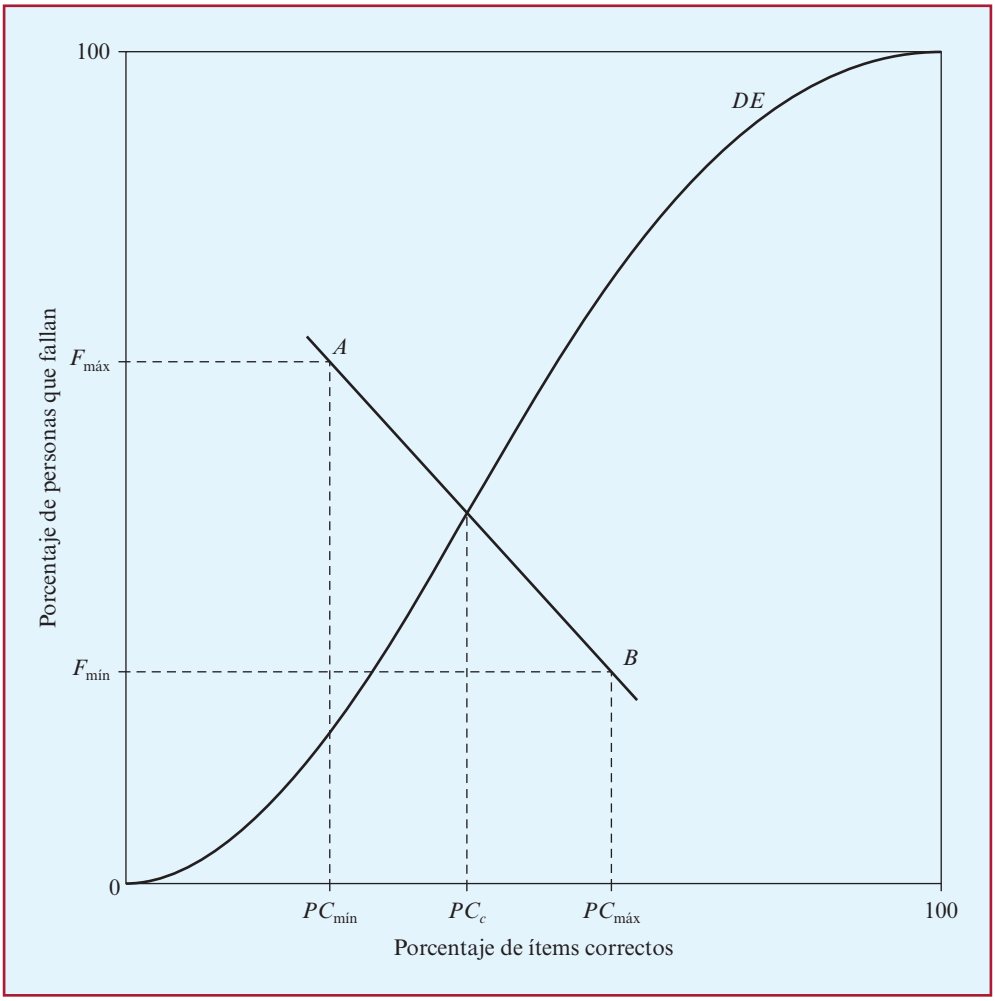


Figura 2.3.—Método de Hofstee.

9.4. Fiabilidad interjueces

Cuando se utilizan jueces o expertos, bien sea para establecer los puntos de corte, tal como se expuso en los apartados anteriores, o en otras muchas situaciones de evaluación, cabe preguntarse en qué medida sus estimaciones resultan fiables, es decir, en qué medida concuerdan sus juicios. Téngase en cuenta que los juicios de los expertos no están exentos de errores de medida, pues, por muy expertos que sean, parafraseando a Terencio, bien se puede decir que nada de lo humano les es ajeno.

Para estimar la fiabilidad interjueces cabe distinguir dos situaciones, en función de cómo se les ha pedido que expresen sus juicios: clasificaciones de las personas en dos o más categorías, o puntuaciones cuantitativas. Veamos cómo proceder en cada uno de los casos.

Clasificaciones

En el caso de que los jueces hayan clasificado a las personas en dos o más categorías, tal como se ha visto en el subepígrafe 8.2, un indicador muy

apropiado es el coeficiente *kappa*. Por ejemplo, en el cuadro adjunto, tomado de dicho subepígrafe, aparecen los datos de 20 personas clasificadas por dos psicólogos clínicos en dos categorías, según consideren que superan o no un determinado nivel de ansiedad. ¿Puede afirmarse que existe una coincidencia diagnóstica entre ambos expertos?

Psicólogo A	**Psicólogo B**		
	Superan	**No superan**	**Total**
Superan	6	1	7
No superan	2	11	13
Total	**8**	**12**	**20**

Como ya se vio en el subepígrafe 8.2, la fórmula del coeficiente *kappa* bien dada por:

$$K = \frac{F_c - F_a}{N - F_a} \qquad [2.63]$$

Aplicando la fórmula, tal como se vio en detalle en el subepígrafe citado:

$$K = \frac{17 - 10{,}6}{20 - 10{,}6} = 0{,}68$$

Este valor indica que el acuerdo entre los dos psicólogos clínicos a la hora de diagnosticar la ansiedad es muy limitado, dado que el valor máximo de *kappa* es 1, que está muy lejos del valor obtenido. Aparte del coeficiente *kappa,* existen otros muchos posibles indicadores para estimar el grado de acuerdo entre clasificaciones. Una buena revisión general puede consultarse en Han y Rudner (2016).

Puntuaciones cuantitativas

Es muy frecuente el caso en el que no se pide a los expertos que clasifiquen a las personas en determinadas categorías, sino que les asignen una puntuación, por ejemplo calificaciones académicas en una escala de cero a diez, u otra cualquiera. En estos casos no tiene sentido utilizar el coeficiente *kappa* y similares, dado que no tenemos categorías; por tanto, hay que usar otro tipo de indicadores para comprobar la fiabilidad interjueces. Vamos a comentar tres muy habituales:

— Correlación de Pearson.
— Coeficiente de concordancia.
— Correlación intraclase.

Correlación de Pearson

El coeficiente de correlación de Pearson (r_{xy}) nos indica el grado de relación lineal entre dos variables; su valor, como es bien sabido, se obtiene dividiendo la covarianza entre las variables correlacionadas por el producto de sus desviaciones típicas:

$$r_{xy} = \frac{\text{cov}(X, Y)}{S_x S_y}$$

Debe quedar muy claro que el coeficiente de correlación de Pearson no es un indicador adecuado para establecer el grado de fiabilidad interjueces. La razón es que puede obtenerse una elevada correlación entre las estimaciones de los jueces, incluso perfecta, y sin embargo encontrarnos ante dos jueces cuyas puntuaciones son muy distintas, por lo que el uso de la correlación de Pearson nos conduciría a error. La razón es muy sencilla: Pearson nos indica la relación lineal entre las puntuaciones de los jueces, de modo que si la correlación es baja, está claro que las puntuaciones de los jueces no convergen, pero si el coeficiente de correlación es elevado, eso no nos garantiza una convergencia adecuada. Por ejemplo, aunque la correlación sea muy elevada, las medias pueden diferir si un juez es más duro que otro en sus apreciaciones. A veces se ha recurrido a complementar Pearson con el cálculo de la diferencia entre las medias, pero incluso podría ocurrir que las medias fuesen iguales y las puntuaciones tuviesen una variabilidad muy diferente. En suma, la correlación de Pearson no es aconsejable para estimar la fiabilidad interjueces.

Veamos un ejemplo. En la tabla adjunta aparecen las puntuaciones en depresión asignadas por dos psicólogas a cinco pacientes. ¿Puede afirmarse que existe buena fiabilidad interjueces?

Nótese que la correlación de Pearson entre las puntuaciones de ambos jueces es 1, es decir, es per-

fecta. Pero ¿es así? ¿son totalmente iguales los diagnósticos de ambas psicólogas? Vemos que la media de las puntuaciones asignadas por la primera psicóloga es 4, y la de la segunda, 8, es decir, la psicóloga 1 puntúa mucho más bajo que la 2. Pero no solo eso: la desviación típica de las puntuaciones otorgadas por la primera es 1,41, mientras que la de la segunda es 2,83, nada menos que el doble. Por tanto, si nos fiásemos del valor de Pearson, cometeríamos un grave error al juzgar el grado de acuerdo entre los diagnósticos de las dos psicólogas. Para evitar estos inconvenientes se han propuesto distintos indicadores, y uno de los más sencillos es el coeficiente de concordancia.

Personas	Psicóloga 1	Psicóloga 2
A	2	4
B	3	6
C	4	8
D	5	10
E	6	12
Media	4	8
Desviación típica	1,41	2,83

Coeficiente de concordancia (C_c)

El coeficiente de concordancia (Lin, 1989) trata de evitar los problemas de la correlación de Pearson para estimar el acuerdo entre expertos. Su fórmula viene dada por la expresión:

$$C_c = \frac{2r_{xy}S_xS_y}{S_x^2 + S_y^2 + (\bar{X} - \bar{Y})^2} \qquad [2.64]$$

Aplicado a los datos de la tabla:

$$C_c = \frac{2(1)(1{,}41)(2{,}83)}{2 + 8 + (4 - 8)^2} = 0{,}31$$

Como se puede observar, el valor obtenido es mucho menor que el ofrecido por la correlación de Pearson, que era 1, lo que indica que efectivamente la fiabilidad interjueces es muy baja para los diagnósticos ofrecidos por las dos psicólogas.

Si hay más de dos jueces, se podría hallar el coeficiente de concordancia para los distintos pares de jueces, pero es más aconsejable utilizar la generalización del coeficiente para *n* jueces, propuesta por Barnhart et al. (2002).

Correlación intraclase

La correlación intraclase (C_{ic}) constituye una alternativa clásica para estimar la fiabilidad interjueces, evitando los problemas que se han mencionado más arriba de la correlación de Pearson. Se trata de un método basado en el análisis de varianza de medidas repetidas, en el que las medidas repetidas son los evaluadores. Se trata, por tanto, de un caso particular del coeficiente de generalizabilidad visto en el subepígrafe 8.4. Por tanto, no existe un solo C_{ic}, sino que su estimación va a depender del diseño de análisis de varianza utilizado. Una guía práctica para elegir el C_{ic} que procede en cada caso puede consultarse en el trabajo de Koo y Li (2016). Tratamientos clásicos detallados sobre la C_{ic} pueden verse en Shrout y Fleiss (1979), McGraw y Wong (1996), Shoukri (2010) o Gwet (2014). Una vez elegido el diseño de ANOVA pertinente, su cálculo está implementado en numerosos programas informáticos, incluido el SPSS. Como los coeficientes anteriores, su valor se encuentra entre 0 y 1. La C_{ic} constituye una alternativa flexible para el cálculo de la fiabilidad interjueces, adaptándose a las distintas situaciones y diseños que se pueden plantear en la práctica. Por supuesto, el uso de la C_{ic} no tiene por qué limitarse al estudio de la fiabilidad interjueces, pues el modelo permite su utilización en cualquier situación en la que se pretenda estimar el grado de acuerdo entre mediciones, por ejemplo ítems, test, como ya se ha visto al tratar de la teoría de la generalizabilidad.

9.5. Comentarios finales

Tras exponer los distintos métodos para establecer los puntos de corte o estándares, seguramente algunos lectores se preguntarán cuál se debe de utilizar o cuál es el mejor. No existe una respuesta unívoca, pero puede consultarse la guía publicada por Berk (1986), en la que revisa y clasifica 38 métodos para establecer estándares, proporcionando tam-

bién consejos prácticos útiles. También puede utilizarse el trabajo de Jaeger (1989). En realidad, todos los métodos funcionan razonablemente bien si se aplican correctamente, aunque lo ideal sería poder aplicar más de uno y así contrastar los resultados. No obstante, la tarea de establecer estándares no es un problema únicamente psicométrico, de modo que aparte del método utilizado debe hacerse uso de todo tipo de información disponible. Es precisamente en la implementación de los métodos en la práctica cuando se suelen cometer los errores, pues, como bien señala Cizek (1996), el peligro está en los detalles. ¿Por qué hay tanto riesgo de aplicar incorrectamente los métodos?: sencillamente porque en todos ellos todo depende de los jueces y hay muchos aspectos de su comportamiento que desconocemos. Empezando por la consistencia o fiabilidad de sus juicios (Plake et al., 1991) y continuando por el número idóneo de jueces en cada caso (Jaeger, 1991), la forma de seleccionarlos y el entrenamiento que debe dárseles (Reid, 1991), la forma ideal de trabajar (Fitzpatrick, 1989): individual o en grupo, etc. Todos ellos son temas a los que se ha dedicado bastante investigación, pero estamos lejos de las respuestas definitivas, entre otras cosas porque son problemas complejos, no circunscritos a la psicometría, que reclaman para su análisis distintas áreas de la psicología y de la sociología. Otro factor clave que subyace a todos los demás es la validez de las opiniones de los jueces (Kane, 1994), que no se debe dar por supuesta, hay que comprobarla. Finalmente, señalar que aquí solo se han abordado métodos surgidos de las necesidades de formas de evaluar más bien clásicas, pero las nuevas orientaciones evaluativas de los noventa están dando lugar a nuevos enfoques para establecer los estándares (Berk, 1996; Clauser y Clyman, 1994; Faggen, 1994; Hambleton y Plake, 1995; Jaeger, 1995; Putnam et al., 1995; Shepard et al., 1993). Tratamientos detallados y actuales pueden consultarse en Hambleton y Pitoniak (2006), Cizek y Bunch (2007), Zieky et al. (2008) o Cizek (2012), y una excelente panorámica del estado actual del campo, en Pitoniak y Cizek (2016). Por su parte, la última versión de los estándares técnicos sobre los test dedica varios apartados al establecimiento de los puntos de corte (AERA, APA, NCME, 2014).

Pitoniak y Cizek (2016) sintetizan el proceso de establecimiento de los puntos de corte en once pasos, que se comentan a continuación:

1. Escoger un método adecuado para establecer los puntos de corte.
2. Elegir los expertos o jueces.
3. Describir las categorías de clasificación que se van a utilizar.
4. Entrenar a los expertos.
5. Definir los conocimientos de las personas que están en el límite de superar los puntos de corte.
6. Recoger las evaluaciones hechas por los expertos.
7. Analizar los datos y dar información a los expertos.
8. Establecer los puntos de corte.
9. Recoger las opiniones de los expertos sobre el desarrollo del proceso.
10. Recoger evidencias de la validez del proceso y preparar la documentación psicométrica.
11. Proporcionar la información correspondiente a los responsables de la evaluación.

Como se puede observar, el proceso general para establecer puntos de corte es amplio y complejo. Aquí nos hemos limitado a ilustrar los métodos más clásicos, correspondientes al primer paso, pero otros muchos aspectos están implicados para llevar la labor a buen fin, como bien ilustran los once pasos citados.

EJERCICIOS

1. Demuestre que en el modelo lineal clásico la covarianza entre los errores y las puntuaciones verdaderas es cero ($\sigma_{ev} = 0$).

2. Demuestre que en el modelo lineal clásico la covarianza entre las puntuaciones empíricas y las verdaderas es igual a la varianza de las verdaderas ($\sigma_{xv} = \sigma_v^2$).

3. Sea el siguiente modelo lineal: $X = V + e_1 + e_2$, donde X y V, como en el modelo lineal clásico, son la puntuación empírica y la verdadera, respectivamente, e_1 es un error de medida debido a las condiciones físicas externas al sujeto y e_2 es un error asociado al estado psicológico del sujeto a la hora de responder al test. Se asume, como parece lógico, que ambos errores están correlacionados y que sus varianzas son iguales; por lo demás, se hacen las mismas asunciones que en el modelo clásico.

1. Exprese σ_X^2 en función de sus componentes.
2. Calcule el valor de la varianza de las puntuaciones empíricas (S_X^2) y el de la covarianza entre empíricas y verdaderas (S_{XV}), sabiendo que la varianza verdadera es 10 ($S_V^2 = 10$), la varianza de los errores 2 ($S_{e_1}^2 = S_{e_2}^2 = 2$) y la correlación entre ambos tipos de error 0,50.
3. Utilizando los datos del apartado anterior, realice los cálculos allí indicados para el modelo:

$$X = V + e_1 - e_2$$

4. A continuación aparecen las puntuaciones empíricas, las verdaderas y los errores correspondientes a cinco sujetos en un test de fluidez verbal y otro de comprensión verbal.

Sujetos	Fluidez verbal			Comprensión verbal		
	X	*V*	*e*	*X*	*V*	*e*
A	5	4	1	6	6	0
B	6	8	–2	8	7	1
C	4	4	0	4	5	–1
D	7	5	2	5	3	2
E	3	0	3	3	5	–2

1. Compruebe si el modelo ($X = V + e$) se cumple para todos los sujetos en ambos test.
2. Ordene a los sujetos en cada test de más favorecido a más perjudicado por los errores aleatorios de medida.
3. Compruebe si para estos datos se cumple el supuesto 2 del modelo en ambos test.
4. Compruebe si para ambos test se cumple: $\sigma_X^2 = \sigma_V^2 + \sigma_e^2$.
5. Conteste, justificando adecuadamente la respuesta, si es correcta o incorrecta la siguiente afirmación: «Si los supuestos del modelo lineal clásico no se cumplen estrictamente para una muestra reducida de sujetos, nunca se cumplirán para la población a la que pertenecen».

5. Dos test presuntamente paralelos se aplicaron a dos muestras independientes de 100 sujetos extraídas de la misma población. La media de las puntuaciones de los sujetos en el primero fue 40, y la del segundo, 36. La varianza de la población es de 9 en ambos casos.

1. Al nivel de confianza del 95%, ¿puede seguir manteniéndose la presunción de paralelismo?
2. A ese mismo nivel de confianza del 95%, ¿qué diferencia máxima entre las medias de ambos test se podría admitir para mantener la presunción de paralelismo?

6. En una muestra de 150 sujetos la media de cierto test fue 70, y su varianza insesgada, 16.

1. ¿Son compatibles estos datos con la hipótesis de que la media del test en la población es 75? NC: 99%.

7. ¿Cuánto vale el coeficiente de fiabilidad de un test cuya varianza verdadera es el 75% de su varianza empírica? ¿Cuál es su índice de fiabilidad?

8. Calcule el coeficiente de fiabilidad de un test en el que la varianza de los errores es el 10% de la varianza empírica.

9. Calcule el coeficiente de fiabilidad de un test en el que la varianza de los errores es el 25% de la varianza verdadera.

10. ¿Cuál es el coeficiente de fiabilidad de un test en el que la varianza de los errores y la varianza verdadera son iguales?

11. Calcule el coeficiente de fiabilidad de un test en el que la varianza verdadera es el 80% de la varianza de los errores.

12. Resuelva los problemas anteriores (7, 8, 9, 10 y 11) sustituyendo en los enunciados el término «varianza» por el de «desviación típica».

13. Demuestre que si se cometiese un error sistemático, el mismo en todos los sujetos, no afectaría al valor del coeficiente de fiabilidad ni al del error típico de medida. Trate de extraer algunas implicaciones de este hecho para la teoría de los test.

14. Represente gráficamente en el eje de ordenadas los valores del error típico de medida correspondientes a los siguientes valores en abscisas del coeficiente de fiabilidad: 0,00, 0,10, 0,20, 0,30, 0,40, 0,50, 0,60, 0,70, 0,80, 0,90, 1,00. Hágalo para valores de 4 y 16 de la varianza empírica. Comente el resultado.

15. Se aplicó un test de rapidez motora a una muestra de 1.000 sujetos, obteniéndose una media de 90, una desviación típica de 10 y un coeficiente de fiabilidad de 0,75. Estime por tres métodos distintos la puntuación verdadera de los sujetos que obtuvieron una empírica de 80. NC: 99%. Compare y comente los resultados obtenidos por los tres métodos.

16. En la desigualdad de Chebychev, ¿cuánto ha de valer K si deseamos hacer afirmaciones con una $p \geqslant 0{,}95$?, ¿y para $p \geqslant 0{,}75$?

17. En una muestra de 2.000 sujetos el coeficiente de fiabilidad de un test fue 0,85, la varianza 64 y la media 60. Al NC del 90%, ¿qué puntuación verdadera se estimará a los sujetos que obtuvieron una empírica de 70?

18. Se aplicó un test a una muestra de 1.500 sujetos. La varianza error resultó ser el 20% de la verdadera; la suma total de las puntuaciones, 60.000, y la varianza empírica, 25. Al NC del 98%, ¿qué puntuación verdadera se estimará a los sujetos que obtuvieron una empírica de 45?

19. La media de una muestra de 100 sujetos en un test fue 20, y la suma de sus puntuaciones empíricas al cuadrado, 50.000. Al NC del 95%, se pronosticó que la puntuación verdadera de un grupo de sujetos estaría entre 40 y 50. ¿Qué puntuación empírica habrán obtenido en el test ese grupo de sujetos?

20. Si la pendiente de la recta de regresión V sobre X es cero, ¿qué pronóstico se hará en V para todo valor de X en puntuaciones directas, diferenciales y típicas?

21. Demuestre que el índice de fiabilidad siempre es mayor o igual que el coeficiente de fiabilidad.

22. Demuestre que $\sigma_X^2 = \sigma_V^2 + \sigma_e^2$.

23. Demuestre que $\rho_{Xe} = \sqrt{1 - \rho_{XX'}}$.

24. Si un test es de velocidad pura y se calcula su coeficiente de fiabilidad por el método de las dos mitades, formadas estas por los ítems pares e impares, respectivamente, ¿cuánto valdría el coeficiente? Razone adecuadamente.

25. Se aplicó un test de 25 ítems a una muestra de 100 sujetos, encontrándose un coeficiente $\alpha = 0{,}42$. Al NC del 95%.

1. ¿Resulta estadísticamente significativo?
2. ¿Son compatibles estos datos con la hipótesis de que el valor de α en la población es 0,53?
3. ¿Entre qué valores se estima que estará el valor de α en la población?
4. Al resolver una de las tres cuestiones anteriores, quedan automáticamente resueltas las otras dos. ¿A cuál nos referimos? Razone adecuadamente.

26. Los errores de medida que afectan a las puntuaciones obtenidas por un grupo de 200 sujetos en un test se distribuyen según la curva normal. La mediana de estos errores es cero, y la suma de sus cuadrados, 288. Calcular:

1. El error típico de medida del test.
2. La desviación típica de las puntuaciones verdaderas, sabiendo que la varianza de las puntuaciones empíricas es 4.
3. La correlación entre las puntuaciones verdaderas y las empíricas.
4. El coeficiente de fiabilidad del test.

27. Calcular la varianza empírica de un test en el supuesto de que la desviación típica de las puntuaciones verdaderas fuese 4, la de los errores 2, y que la correlación entre las puntuaciones verdaderas y los errores de medida fuese 0,50.

28. Se aplicó un test a una muestra de 85 sujetos, obteniéndose que la suma de sus puntuaciones diferenciales al cuadrado fue 1.360, y la varianza de los errores, 9. ¿Cuánto vale el coeficiente de fiabilidad de dicho test?

29. En un test de inteligencia espacial, una muestra de 200 sujetos obtuvo una media de 50, y una desviación típica de los errores de medida igual a 2, lo que supone un 20% de la desviación típica de las puntuaciones verdaderas.

1. ¿Qué puntuación verdadera diferencial corresponderá a los sujetos que obtuvieron una puntuación empírica directa de 70? (NC: 95%).
2. A los sujetos con una determinada puntuación empírica se les pronosticó que su puntuación verdadera estaría entre 10 y 20. ¿A qué nivel de confianza se habrá hecho?
3. Al NC del 90%, ¿qué error máximo estamos dispuestos a admitir que afecta a las puntuaciones?

30. Se aplicó un test de fluidez verbal a un grupo de 100 sujetos. A los que tenían una puntuación empírica diferencial de 4 puntos se les estimó (NC del 95%) que su puntuación diferencial verdadera estaría entre 6,92 y –0,92. Sabiendo que la media del grupo en el test fue de 8 puntos, calcular:

1. El coeficiente de fiabilidad.
2. El error típico de medida.
3. La pendiente de la recta de regresión de V/X en puntuaciones diferenciales.
4. La puntuación verdadera diferencial que se estimará a los sujetos que obtuvieron en el test una puntuación empírica directa de 10 puntos (NC del 96%).

31. Se aplicó un test a una muestra aleatoria de 1.000 estudiantes de la Universidad de Oviedo. La media de sus puntuaciones en el test fue 25; la varianza de las puntuaciones verdaderas, 64, y la de los errores, 9. ¿Cuál es el intervalo confidencial en el que se puede afirmar que se encontrará la puntuación verdadera correspondiente a una empírica directa de 33 puntos? (NC del 95%).

32. Se aplicó un test de inteligencia general a una muestra de 100 sujetos, obteniéndose una media de 40 puntos y una varianza de 25. El índice de fiabilidad del test para esa muestra fue de 0,80. ¿Entre qué valores se encontrará la puntuación diferencial verdadera de los sujetos que obtuvieron en el test una puntuación empírica directa de 50 puntos? (NC del 95%).

33. Se desea pronosticar las puntuaciones verdaderas de un test a partir de las empíricas. La pendiente de la recta de regresión de V/X en puntuaciones diferenciales es de 0,81. ¿Cuál será la puntuación típica verdadera pronosticada a los sujetos que obtuvieron en el test una puntuación típica empírica de 0,50?

34. La media de una muestra de alumnos de tercer curso de psicología en un test de destreza manual fue de 50 puntos, y la desviación típica, de 15. La desviación típica de las puntuaciones verdaderas resultó ser el 85% de la de las empíricas.

1. Calcular el coeficiente de fiabilidad del test.
2. ¿Entre qué valores se estima que se encontrará la puntuación directa verdadera de los alumnos que obtuvieron en el test una puntuación empírica directa de 55 puntos? (NC del 99%).
3. Si el test de destreza manual se aplicase a una muestra con una varianza de 81 puntos, ¿cuál sería su coeficiente de fiabilidad?

35. Una muestra de alumnos obtuvo en un test de fluidez verbal una media de 20 y una desviación típica de 5, siendo la desviación típica de los errores de medida el 30% de la desviación típica de las puntuaciones empíricas.

1. Calcular el coeficiente e índice de fiabilidad.
2. Calcular la correlación entre las puntuaciones empíricas y los errores de medida.

36. En una muestra aleatoria de sujetos, un test de inteligencia verbal tiene una media de 10, una desviación típica de 3 y una varianza error de 0,81. Calcular:

1. El coeficiente de fiabilidad del test.
2. El error típico de medida del test.
3. La varianza de las puntuaciones verdaderas.
4. El coeficiente de fiabilidad que tendría ese mismo test en una muestra con una varianza de 2.

37. Se aplicó un test a una muestra de 1.000 sujetos. El coeficiente de fiabilidad fue de 0,64 y la desviación típica, de 2. ¿Cuál sería el coeficiente de fiabilidad del test si se aplicase a una muestra de 500 sujetos con una varianza de 16?

38. Para la selección de personal de una empresa se aplicó un test de rapidez perceptiva a un grupo de aspirantes, obteniéndose una desviación típica de 25. Solo se admitió al 10% de los aspirantes, en cuyo grupo el índice de fiabilidad del test resultó ser de 0,90, y la desviación típica, de 5. ¿Cuál se estima que sería el coeficiente de fiabilidad del test si se hubiese calculado en el grupo de aspirantes y no en el de admitidos?

39. Si la correlación entre las dos mitades paralelas de un test es de 0,85, ¿cuál es su índice de fiabilidad?

40. Aplicada una escala de dogmatismo a una muestra de sujetos, se obtuvo una varianza de las puntuaciones empíricas de 10 y una varianza error de 0,90. Calcular:

1. Su coeficiente de fiabilidad.
2. Su índice de fiabilidad.
3. El coeficiente de fiabilidad que tendría si se duplicase su número de ítems.

41. La covarianza entre las puntuaciones pares e impares de un test fue de 14,4 y la varianza total del test 60,8. Teniendo en cuenta que la varianza de las dos mitades (par e impar) fue la misma, ¿cuánto valdrá el coeficiente de fiabilidad del test?

42. Sea:

$$Z = X + Y$$
$$W = X - Y$$
$$S_Z^2 = 380$$
$$S_W^2 = 20$$

Calcular el índice de fiabilidad del test Z, teniendo en cuenta que X e Y son dos formas paralelas.

43. Sean X e Y dos test paralelos. La varianza de las puntuaciones $D = X - Y$ vale 73,28 y la varianza de $S = X + Y$ es igual a 841,56.

1. Calcular el coeficiente de fiabilidad de S.
2. Calcular el coeficiente de fiabilidad de X e Y.

44. Una muestra de 1.000 niños obtuvo en una prueba de pensamiento divergente una media de 40 puntos y una desviación típica de 5, siendo la desviación típica de las puntuaciones verdaderas el 90% de la desviación típica de las puntuaciones empíricas.

1. Calcular el coeficiente de fiabilidad del test.
2. ¿Entre qué valores se encontrará la puntuación directa verdadera de los niños que obtuvieron en el test una puntuación directa empírica de 45 puntos? (NC: 99%).
3. Calcular la fiabilidad del test si se le añadiesen 20 ítems paralelos a los 80 que poseía inicialmente.
4. Si el test inicial (80 ítems) se aplica a una muestra con una varianza de 49 puntos, ¿cuál sería su fiabilidad para este nuevo grupo?

45. Si se desea construir un test con un coeficiente de fiabilidad de 0,90 y para ello se dispone de dos test apropiados, uno con 5 ítems y un coeficiente de fiabilidad de 0,50, y otro con 10 ítems y un coeficiente de fiabilidad de 0,60, ¿cuál de ellos elegiríamos para alcanzar dicho coeficiente de fiabilidad con el menor número de ítems?

46. La desviación típica de los errores en un test de 150 ítems es el 40% de la desviación típica de las puntuaciones empíricas.

1. ¿Cuál será el coeficiente de fiabilidad del test si se le suprimen 60 de sus ítems?
2. ¿Cuántos de los 150 ítems originales habría que retener para lograr una fiabilidad de 0,70?

47. Para investigar el área verbal, tanto su aspecto productivo (fluidez) como comprensivo (comprensión), se aplicó un test de fluidez verbal (FV) y otro de comprensión verbal (CV) a una muestra de 500 sujetos. En el caso de la FV, la varianza verdadera fue el 85% de la empírica, y en el de la CV, el 90%.

1. Teniendo en cuenta que el coeficiente de fiabilidad de la CV se calculó por el método de las dos mitades, ¿cuál fue la correlación entre ambas mitades?
2. Al test de CV que constaba originalmente de 120 ítems se le suprimieron 40 por considerarlo muy fatigoso para los sujetos; ¿cuánto valdrá su fiabilidad tras acortarlo?
3. ¿Cuántos ítems habrá que suprimirle al test de FV, que constaba de 100, si nos conformamos con una fiabilidad de 0,80?
4. Si ambos test tuviesen el mismo número de ítems, ¿cuál sería más fiable?
5. A la hora de vender el test de FV a una institución, esta impone dos condiciones: *a*) que el test tenga únicamente 60 ítems y *b*) que su fiabilidad no sea inferior a 0,82 para muestras con una desviación típica en el test no superior a 15. ¿En qué condiciones cumple el test de FV los requisitos exigidos? (S_{FV} = 10).

48. Un test que consta de 40 ítems paralelos tiene una varianza global de 25, y el coeficiente de fiabilidad de cada ítem es 0,12.

1. Calcular el coeficiente de fiabilidad del test.
2. Calcular la correlación entre las puntuaciones empíricas y los errores de medida.
3. Al NC del 95%, ¿qué error máximo afectará a las puntuaciones verdaderas pronosticadas?
4. Si la fiabilidad de cada ítem fuese 0,15, ¿qué proporción representaría la varianza verdadera del test respecto de la empírica?

49. A una muestra de 100 sujetos se le aplicó un test de independencia de campo. La suma de los cuadrados de los errores de medida fue 256, distribuyéndose dichos errores aleatorios según la curva normal, con media cero.

1. Calcular el error típico de medida del test.
2. Teniendo en cuenta que la varianza de las puntuaciones empíricas en el test fue 4, calcular la desviación típica de las puntuaciones verdaderas.
3. Calcular el coeficiente de fiabilidad del test.
4. ¿Cuánto valdrá la correlación entre las puntuaciones verdaderas y las empíricas?
5. Administrado el mismo test a otra muestra de sujetos, se obtuvo una desviación típica de 8. ¿Qué coeficiente de fiabilidad cabe esperar en esta segunda muestra?

50. Un test de inteligencia general se aplicó a una muestra de 500 sujetos, obteniéndose una varianza de las puntuaciones empíricas de 882. Dividido el test en dos mitades paralelas, *B* y *C*, se encontró entre ellas una correlación de 0,96.

1. Calcular el coeficiente de fiabilidad del test.
2. Sabiendo que la media del test fue 50, ¿qué puntuación empírica directa habrán obtenido los sujetos a los que se les ha pronosticado una puntuación directa verdadera comprendida entre 62,98 y 46,82?
3. Calcular la varianza empírica de *B*.
4. Calcular la varianza verdadera de *C*.
5. ¿Cuál sería el coeficiente de fiabilidad del test si se aplicase a un grupo de sujetos cuya varianza en dicho test fuese de 64?

51. La media de una muestra de 100 sujetos en una escala de neuroticismo de 80 ítems es 20, la desviación típica de las puntuaciones verdaderas es el 90% de la de las empíricas y la media de los errores cuadráticos de medida vale 9. Se sabe además que, al nivel de confianza del 95%, se pronosticó que la puntuación verdadera correspondiente a cierta empírica estaría entre 9 y 21.

1. Calcular el índice y el coeficiente de fiabilidad.

2. Calcular el error típico de medida.
3. Calcular la desviación típica de las puntuaciones empíricas.
4. Calcular la puntuación empírica directa, diferencial y típica cuya verdadera se estima que estará entre 9 y 21.
5. ¿Cuál sería el coeficiente de fiabilidad de la escala si el número de ítems se redujese a la mitad?
6. ¿Cuántos ítems tendría que tener la escala para que el coeficiente de fiabilidad fuese de 0,94?
7. ¿Cuál sería la fiabilidad de la escala en otra muestra con desviación típica doble?

52. La varianza de las diferencias entre las puntuaciones pares e impares de un test es 36, y la varianza empírica del test total, 100. ¿Cuál es el índice de fiabilidad del test?

53. La media de un test de 100 ítems en una muestra de 400 sujetos fue 45, la desviación típica 14, y la covarianza entre las dos mitades paralelas, formadas por los ítems pares e impares, respectivamente, resultó ser 44,1. Calcular:

1. El coeficiente de fiabilidad del test.
2. La correlación entre las dos mitades del test.
3. Las varianzas de la mitad par e impar.

54. Una escala formada por 10 ítems de razonamiento espacial se aplicó a una muestra de ocho sujetos. Los resultados aparecen en la tabla adjunta, donde el 1 significa que el sujeto superó el ítem y el 0 que lo falló.

1. Calcular la fiabilidad por el método de Rulon.
2. Calcular la fiabilidad por el método de Guttman-Flanagan.
3. Calcular la fiabilidad por el método de las dos mitades.
4. Calcular α, KR_{20} y KR_{21} y comparar los resultados.

Sujetos	Ítems									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	1	0	1	1	1	1	1	1	0	1
B	1	1	1	1	1	1	1	1	1	0
C	1	1	1	1	1	1	0	1	1	0
D	1	1	1	1	1	1	1	1	1	1
E	1	1	1	1	1	1	1	1	1	1
F	1	1	1	1	1	1	1	0	0	0
G	1	1	1	1	1	1	1	0	1	0
H	1	1	1	1	1	0	1	0	0	0

55. Se aplicó un test de 10 ítems a una muestra de sujetos, obteniéndose los siguientes resultados: media 52,625, varianza 345,11, varianza de la mitad formada por los ítems pares 87,03, varianza de los impares 88,87. Calcular:

1. El coeficiente de fiabilidad del test.
2. El índice de fiabilidad del test.
3. La desviación típica de los errores de medida del test.
4. La correlación entre las puntuaciones empíricas y los errores de medida.
5. La varianza de las puntuaciones verdaderas.
6. El intervalo confidencial en el que se puede afirmar, al NC del 95%, que se encuentra la puntuación típica verdadera de los sujetos que obtuvieron en el test una puntuación empírica directa de 60 puntos.
7. Si la suma de las varianzas de los 10 ítems fuese 47,6, ¿cuánto valdría el coeficiente α?

8. Calcular el valor de la covarianza media entre los ítems del test.

56. Un test de tres ítems se aplicó a un grupo de cuatro sujetos, obteniéndose los datos de la tabla adjunta.

Sujetos	Ítems		
	1	2	3
A	1	0	1
B	0	1	1
C	1	0	1
D	0	0	0

Calcule el coeficiente de fiabilidad por el método que considere más adecuado. Justifique la elección y comente el resultado.

57. Una batería de rendimiento escolar consta de tres pruebas: geografía, lengua y matemáticas. Aplicada la batería a una amplia muestra de escolares asturianos, se obtuvieron varianzas empíricas de 10, 12 y 16, respectivamente, covarianzas de $S_{gl} = 2$, $S_{gm} = 1{,}5$ y $S_{lm} = 2{,}5$, y coeficiente de fiabilidad de 0,80, 0,90 y 0,90, respectivamente.

1. Calcular el coeficiente de fiabilidad de la batería.
2. Calcule el coeficiente α de la batería y compárelo con el resultado del apartado anterior. Trate de explicar la razón de las discrepancias entre ambos, si las hubiere.
3. ¿Cuál sería la fiabilidad de la batería para un extraño colegio que ponderase las puntuaciones en matemáticas con un peso de 2 y las de geografía y lengua con –1?

58. En una investigación sobre comprensión verbal, se aplicó a una muestra de cinco sujetos un test compuesto por los subtest *A* y *B*. Los resultados aparecen en la tabla adjunta.

1. Calcule el coeficiente de fiabilidad del subtest *A* por el método de Rulon.
2. Calcule la fiabilidad del subtest *B* en función de la consistencia interna de sus ítems.
3. Calcule la covarianza media de los ítems de *B*.
4. Calcule la fiabilidad del test global.

Sujetos	Subtest *A*				Subtest *B*		
	Ítems				Ítems		
	1	2	3	4	1	2	3
A	0	0	0	0	1	1	1
B	1	0	0	1	1	1	1
C	0	1	1	1	1	0	0
D	0	0	0	1	0	0	0
E	1	1	1	1	0	1	0

59. Imagine que está trabajando con el test de destreza manual «Tablero de Clavijas de Purdue», consistente en que los sujetos tienen que introducir el mayor número posible de clavijas en un tablero cuyas ranuras se ajustan con precisión a las clavijas. La puntuación asignada a los sujetos es el número de clavijas colocadas.

1. Describa concisa y claramente los pasos que seguiría para obtener el coeficiente de fiabilidad del test. Justificar el tipo de coeficiente elegido.
2. En el supuesto de que el test resultase con un elevado coeficiente de fiabilidad, ¿qué nos garantizaría exactamente acerca de la medida de la destreza manual?
3. Si los clientes estuviesen interesados en conocer el coeficiente α del test, ¿cómo lo calcularía? Razone adecuadamente la respuesta.

60. Se aplicó un test de cinco ítems a una muestra de cuatro sujetos obteniéndose los datos de la tabla adjunta, donde 1 significa acierto y 0 fallo.

Sujetos	Ítems				
	1	2	3	4	5
S_1	1	1	1	1	0
S_2	1	0	1	1	1
S_3	1	0	1	0	0
S_4	1	0	0	0	0

1. Calcule el coeficiente α y comente el resultado.
2. Al nivel de confianza del 95%, ¿puede afirmarse que α es estadísticamente significativo?
3. Según estos datos, y al nivel de confianza del 95%, ¿cuál se estima que será el valor de α en la población?
4. Calcule un estimador insesgado de α y trate de dar una explicación del extraño resultado encontrado, si es que lo es.

61. En la tabla adjunta aparecen las puntuaciones obtenidas por cinco sujetos en dos formas paralelas de un test X que consta de 10 ítems.

X	X'
4	5
3	2
0	1
2	3
1	1

1. Calcule el coeficiente de fiabilidad del test. Comente el resultado.
2. ¿Cuántos ítems habría que añadir al test para que alcanzase una fiabilidad de 0,90?
3. Calcule los errores de sustitución para los cinco sujetos.
4. Calcule los errores de predicción para los cinco sujetos.

62. En la matriz adjunta aparecen las puntuaciones de cinco sujetos en dos test paralelos A y A'.

Sujetos	Subtest A				Subtest A'			
	Ítems				Ítems			
	1	2	3	4	1	2	3	4
A	1	0	0	0	0	0	0	0
B	1	1	0	0	1	1	0	0
C	1	1	1	0	1	1	1	1
D	1	1	1	1	1	1	1	0
E	0	0	0	0	1	0	0	0

1. Calcule el coeficiente de fiabilidad del test A y comente el resultado.
2. ¿Qué puntuación verdadera se estima que obtendrán en el test A aquellos sujetos que obtengan una empírica directa de 3? Nivel de confianza: 92%.
3. Calcule el coeficiente α de cada una de las dos formas paralelas (A y A') y compruebe si la diferencia entre ambos coeficientes resulta estadísticamente significativa. NC del 95%.
4. ¿Cuál sería el coeficiente de fiabilidad del test formado por el A y el A' juntos?
5. Compruebe que, efectivamente, tal como se dice en el enunciado, el test A y el A' son paralelos.

63. Se aplicó un test de inteligencia verbal de 40 ítems a una muestra de cinco sujetos: A, B, C, D, E. Las puntuaciones de cada sujeto en la mitad par e impar aparecen entre paréntesis: $A(8,10)$, $B(6,4)$, $C(0,2)$, $D(4,6)$, $E(2,2)$. El primer número del paréntesis corresponde a la puntuación del sujeto en la mitad par, y el segundo, a la mitad impar.

1. Calcular el coeficiente de fiabilidad del test por el método de las dos mitades e interpretar el resultado.
2. Calcular el error típico de medida.
3. Elaborar la recta de regresión en puntuaciones directas, diferenciales y típicas para pronosticar las puntuaciones verdaderas a partir de las empíricas.
4. Al NC del 98%, ¿qué puntuación verdadera se estimará a los sujetos que obtuvieron una empírica de 5?
5. ¿Cuántos ítems habría que retener si nos conformásemos con que el test tuviese un coeficiente de fiabilidad de 0,80?
6. Haga una valoración razonada de la consistencia interna del test de inteligencia verbal.
7. Si el test de inteligencia verbal se aplicase a una muestra de sujetos cuya varianza en el test fuese l00, ¿cuál sería el coeficiente de fiabilidad en esas condiciones?
8. Calcular el coeficiente de fiabilidad del test original de inteligencia verbal por el método de Rulon.

9. A la vista de los datos del enunciado general, razone adecuadamente acerca de si el test de inteligencia verbal es de velocidad, de potencia o mixto.
10. ¿Qué puntuaciones tendrían que haber sacado los cinco sujetos en la segunda mitad del test para que la correlación con la primera fuese perfecta? Justifique la respuesta adecuadamente.

64. Se aplicaron dos test de razonamiento numérico a una muestra de cinco sujetos, obteniéndose los resultados que aparecen en la tabla adjunta. El primer test constaba de tres ítems, y el segundo, de cuatro.

Sujetos	Test 1			Test 2			
	1	2	3	1	2	3	4
A	1	1	1	1	1	1	1
B	1	1	0	1	1	1	0
C	1	1	0	1	1	0	0
D	0	0	0	1	0	0	0
E	1	0	0	0	0	0	0

1. Al NC del 95%, ¿puede afirmarse que la consistencia interna del segundo test es superior a la del primero?
2. Calcular el coeficiente α de la batería formada por ambos test.
3. Al NC del 95%, ¿entre qué valores se estima que estará el valor paramétrico del coeficiente α de la batería?
4. Al NC del 95%, ¿puede afirmarse que el coeficiente α de la batería es estadísticamente significativo?
5. Si las puntuaciones de los sujetos en el primer test se ponderasen con un peso de 3, y las del segundo con un peso de 2, calcular los efectos de esas ponderaciones sobre: 1) la varianza de la batería formada por los dos test; 2) la correlación entre los dos test, y 3) la covarianza entre los dos test.

65. En la tabla adjunta aparecen los datos obtenidos al aplicar un test de inteligencia espacial a una muestra de cinco sujetos en dos ocasiones, con un período intermedio de una semana.

Sujetos	Test	
	1.ª	2.ª
A	6	8
B	8	6
C	5	5
D	4	2
E	2	4

1. Calcule el coeficiente de fiabilidad del test. Comente el resultado.
2. Al nivel de confianza del 92%, ¿entre qué valores se estima que se encontrará la puntuación verdadera en el test de los sujetos que hayan obtenido una puntuación empírica de 7 puntos?
3. Si se desean hacer los pronósticos sobre las puntuaciones verdaderas con un error máximo que no exceda la unidad, ¿a qué nivel de confianza deberíamos trabajar? Comente el resultado.
4. Teniendo en cuenta que la covarianza media entre los cuatro ítems del test fue de 0,10, ¿cuál es el coeficiente α del test?

66. En el manual editado para los usuarios de cierto test, el editor recomienda que las puntuaciones de los sujetos se les comuniquen en forma de una banda comprendida entre un error típico de medida por encima y por debajo de la puntuación empírica ($X \pm \sigma_e$).

¿A qué nivel de confianza se están comunicando las puntuaciones a los usuarios?

67. Se aplicó un test de razonamiento verbal de 20 ítems a una muestra de 16 personas. Los resultados aparecen en la tabla adjunta, en la que 1 significa acierto en el ítem y 0 error.

1. Calcular el coeficiente de fiabilidad del test mediante la fórmula de Rulon.
2. Calcular el error típico de medida del test.
3. Calcular la puntuación verdadera que se estimará a una persona que obtuvo en el test una empírica de 12. Utilizar el error típico de medida para la estimación. NC 95%.
4. Divididas las puntuaciones del test en cuatro niveles, calcular el error típico de medida para cada uno de ellos.

Sujetos	Ítems																			
A	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	1	1	1	1	0	0	0	0	1	1	0	1	1	1	0	1	0	0	0	0
D	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	0
E	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
F	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
H	0	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0
I	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0
J	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
K	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
L	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0
M	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
N	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Ñ	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
O	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5. Utilizando un error típico de medida del nivel correspondiente, estimar la puntuación verdadera de las personas con una puntuación empírica de 12 puntos en el test. NC 95%.
6. Compare los resultados obtenidos en los apartados 3 y 5 y explique la razón de las diferencias, si las hubiere.
7. ¿A qué nivel mide el test con menor precisión? Razone la respuesta.
8. Se estableció un punto de corte en la puntuación 16 para clasificar a los examinados en aprobados y suspensos. Se considera que un intervalo vital para una clasificación apropiada viene dado por ±2 unidades en torno al punto de corte. ¿Cuál es el error típico de medida para dicho intervalo?
9. ¿Cuál es la probabilidad de que una persona con una puntuación verdadera de 15 puntos en el test supere el punto de corte establecido en 16 puntos? Se asume que los errores de medida se distribuyen según la curva normal en torno a la puntuación verdadera.
10. Bajo los mismos supuestos del apartado anterior, ¿cuál es la probabilidad de que una persona con una puntuación verdadera de 18 puntos no supere el punto de corte establecido en 16 puntos?

68. Siendo e el error del test global y e_1 y e_2 los errores de cada una de las mitades, demostrar que la varianza de los errores globales del test es igual a la suma de las varianzas de los errores de cada una de las dos mitades: $\sigma_e^2 = \sigma^2(e_1 - e_2) = \sigma_{e_1}^2 + \sigma_{e_2}^2$.

69. En la tabla 2.8 aparecen las puntuaciones de 20 personas en dos formas paralelas de un test referido al criterio. Dichas formas se utilizan para clasificar a las personas en tres categorías: baja (puntuaciones 0-3), media (puntuaciones 4-7) y alta (puntuaciones 8-10). Elabore la tabla correspondiente y calcule los coeficientes de fiabilidad p_o y k.

70. Calcule el coeficiente de fiabilidad clásico de formas paralelas para los datos de la tabla 2.8. Compare su valor con el de los coeficientes p_o y *kappa* y comente el resultado.

71. Calcule la correlación de Pearson (coeficiente ϕ) para los datos de la tabla 2.9. Compare su valor con el de los coeficientes p_o, *kappa* y $r_{xx'}$ del ejercicio anterior. Comente el resultado.

72. Calcule el valor del coeficiente de Livingston cuando se dispone de dos formas paralelas. Utilice los datos de la tabla 2.8.

73. En la tabla adjunta aparecen subrayadas las alternativas erróneas que tres jueces consideran que serían detectadas por una persona con los co-

nocimientos mínimos requeridos para superar la materia. Cada ítem tiene cinco alternativas, apareciendo con un asterisco la correcta.

1. Calcule el valor esperado asignado al test por cada juez.
2. Establezca el punto de corte según el método de Nedelsky, sin corregir los efectos del azar y corrigiéndolos.

Ítems	Juez A	Juez B	Juez C
1	*a*bcde*	*a* bcde*	*a*bcde*
2	*ab*cde*	*ab*cde*	*ab*cde*
3	*ab*cde*	*ab*cde*	*ab*cde*
4	*abcd*e*	*abcd*e*	*abcd*e*
5	*abc*de*	*abc*de*	*abc*de*

74. En la tabla adjunta aparecen las probabilidades asignadas por cuatro jueces de que los ítems de un test sean superados por personas con unos conocimientos mínimos exigibles. Todos los ítems son de respuesta abierta.

Ítems	Juez A	Juez B	Juez C	Juez D
1	0,8	1,0	0,9	0,8
2	0,9	0,8	1,0	0,7
3	0,6	0,5	0,5	0,4
4	0,5	0,5	0,4	0,3
5	0,4	0,4	0,4	0,4
6	0,3	0,2	0,3	0,2

1. Obtenga mediante el método de Angoff los puntos de corte correspondientes a cada juez.
2. Calcule el punto de corte del test, teniendo en cuenta las opiniones de todos los jueces. Justifique el estadístico elegido.
3. Según los datos de la tabla, ¿cuál de los jueces considera más fácil el test? Justifique la respuesta.
4. Teniendo en cuenta la opinión de los cuatro jueces, ¿cuál es el ítem más difícil del test? Justifique la respuesta.

75. Tras analizarlos conjuntamente, los cuatro jueces del ejercicio 74 llegaron a un consenso para clasificar los seis ítems, siguiendo el método de Ebel, del siguiente modo:

	Dificultad	Relevancia	Superan
Ítem 1	Media	Fundamental	70%
Ítem 2	Difícil	Fundamental	40%
Ítems 3 y 4	Media	Importante	80%
Ítem 5	Fácil	Aceptable	50%
Ítem 6	Difícil	Aceptable	30%

1. Elabore la tabla de 12 categorías sugerida por Ebel y calcule el punto de corte del test por el método propuesto por este mismo autor.
2. Observe los juicios emitidos individualmente por los jueces en el ejercicio 74 y los generados por consenso por esos mismos jueces en este ejercicio. Analice las diferencias. Acudiendo a sus conocimientos de psicología social, trate de profundizar en el estudio de la incidencia del número de jueces y de la forma de elaboración de los juicios (individual/grupo) sobre los juicios emitidos.

SOLUCIONES

3.1. $\sigma_X^2 = \sigma_V^2 + 2\sigma_e^2 + 2\rho_{e_1e_2}\sigma_e^2$
2. $S_X^2 = 16$; $S_{XV} = 10$
3. $S_X^2 = 12$; $S_{XV} = 10$

4.1. Sí
2. *FV*: *E*, *D*, *A*, *C*, *B*
   CV: *D*, *B*, *A*, *C*, *E*
3. *FV*: $r_{Ve} = -0{,}85$. No se cumple
   CV: $r_{Ve} = -0{,}21$. No se cumple
4. *FV*: $S_X^2 = 1{,}99$; $S_V^2 = 6{,}56$
   $S_e^2 = 2{,}96$. No se cumple
   CV: $S_X^2 = 2{,}96$; $S_V^2 = 1{,}76$
   $S_e^2 = 1{,}99$. No se cumple
5. Incorrecto

5.1. No: $Z = 9{,}43 > 1{,}96$
2. 0,83

6.1. No: $T = -15{,}31 < -2.617$

7. 0,75; 0,866
8. 0,90
9. 0,80
10. 0,50
11. 0,45
12. 0,5625; 0,75; 0,99; 0,94; 0,50; 0,39
14.

$r_{xx'}$	S_{e_1}	S_{e_2}
0,00	2,000	4,000
0,10	1,897	3,795
0,20	1,788	3,577
0,30	1,673	3,346
0,40	1,549	3,098
0,50	1,414	2,828
0,60	1,265	2,529
0,70	1,095	2,191
0,80	0,894	1,788
0,90	0,632	1,265
1,00	0,000	0,000

15. $30 \leqslant V \leqslant 130$
$67{,}10 \leqslant V \leqslant 92{,}90$
$71{,}33 \leqslant V \leqslant 93{,}67$
16. $\sqrt{20}$; 2,00
17. $63{,}82 \leqslant V \leqslant 73{,}18$
18. $39{,}77 \leqslant V \leqslant 48{,}53$
19. 46,88
20. La media, 0,0
24. 1
25.1. Sí: $F = 1{,}724 > 1{,}27$
2. Sí: $F = 0{,}81 < 1{,}27$
3. $0{,}2634 \leqslant \alpha \leqslant 0{,}5592$
4. 3.ª
26.1. 1,20
2. 1,60
3. 0,80
4. 0,64
27. 28
28. 0,4375
29.1. $15{,}30 \leqslant V \leqslant 23{,}10$
2. 98,8 %
3. 3,26
30.1. 0,75
2. 2,31
3. 0,75
4. $-2{,}60 \leqslant V \leqslant 5{,}60$
31. $26{,}506 \leqslant V \leqslant 37{,}526$
32. $1{,}696 \leqslant V \leqslant 11{,}104$
33. 0,45
34.1. 0,7225
2. $36{,}27 \leqslant V \leqslant 70{,}95$
3. 0,23
35.1. 0,91; 0,95
2. 0,30
36.1. 0,91
2. 0,90
3. 8,19
4. 0,595
37. 0,91
38. 0,99
39. 0,958
40.1. 0,91
2. 0,95
3. 0,953
41. 0,947
42.1. 0,97
43.1. 0,91
2. 0,84
44.1. 0,81
2. $38{,}993 \leqslant V \leqslant 49{,}107$
3. 0,84
4. 0,90
45. $N_1 = 45$, $N_2 = 60$
Se elige el primero
46.1. 0,759
2. 67
47.1. 0,82
2. 0,85
3. 29
4. CV
5. $11{,}30 \leqslant S_X \leqslant 15$
48.1. 0,845
2. 0,39
3. 3,53
4. 0,87
49.1. 1,60
2. 1,20
3. 0,36
4. 0,60
5. 0,96
50.1. 0,979
2. 55
3. 225
4. 216
5. 0,71
51.1. 0,90; 0,81
2. 3
3. 6,88
4. 13,83; –6,17; –0,897

5. 0,68
6. 294
7. 0,95

52. 0,80

53.1. 0,90
2. 0,82
3. 53,78

54.1. 0,59
2. 0,59
3. 0,68
4. 0,43; 0,43; 0,16

55.1. 0,98
2. 0,989
3. 2,62
4. 0,14
5. 338,24
6. $0,12 \leqslant Z_V \leqslant 0,66$
7. 0,95
8. 3,30

56.1. 0,25

57.1. 0,904
2. 0,36
3. 0,87

58.1. 0,88
2. 0,75
3. 0,12
4. 0,73

59.1. Test-retest, formas paralelas
2. Error de medida bajo
3. No calculable

60.1. 0,648
2. $F = 2,84$; $F_{0,975(3,12)} = 4,47$; $F_{0,025(3,12)} = 0,069$; No
3. $(-0,57 \leqslant \alpha \leqslant 0,975)$
4. 0,88

61.1. $r_{xx'} = 0,85$
2. 6
3. –1, +1, –1, –1, 0
4. –0,08, 1,32, –0,88, –0,48, 0,12

62.1. 0,80
2. $(1,82 \leqslant V \leqslant 3,78)$
3. $\alpha_A = 0,80$, $\alpha_{A'} = 0,80$
4. 0,89
5. Sí: medias y varianzas iguales

63. 1. 0,92
2. 1,58
3. $V' = 0,92X + 0,704$
   $V' = 0,92x$
   $Z_{V'} = 0,959Z_X$
4. $1,764 \leqslant V \leqslant 8,844$
5. 14
6. $r_{pi} = 0,85$
7. 0,975
8. 0,92
9. Mixto/potencia
10. Cualquier transformación lineal de la primera

64.1. No: 0,685 < 2,353
2. 0,875
3. $0,577 \leqslant \alpha \leqslant 0,985$
4. Sí: $(0,577 \leqslant \alpha \leqslant 0,985)$, no incluye $\alpha = 0$
5. 1) 31,677; 2) 0,83; 3) 7,179

65.1. 0,60
2. $4,49 \leqslant V \leqslant 7,91$
3. 69,22%
4. $\alpha = 0,40$

66. 68,26%

67. 1. 0,93
2. 1,45
3. $9,16 \leqslant V \leqslant 14,84$
4. 0,71; 0,87; 0,71; 1,87
5. $10,61 \leqslant V \leqslant 13,39$
6. $\sigma_{e_3} < \sigma_e$
7. 4
8. 1,78
9. 0,29
10. 0,13

69. $p_o = 0,65$; $K = 0,45$

70. $r_{xy} = 0,75$

71. $\phi = 0,68$

72. 0,83

73.1. 2,33; 2,08; 1,83
2. 2,08; 1,35

74.1. $A = 3,5$; $B = 3,4$; $C = 3,5$; $D = 2,8$
2. $PC = 3,3$
3. Jueces A y C
4. Ítem 6

75.1. $PC = 3,5$

3 Validez

1. CONCEPTO

Un test constituye una muestra de conducta de una persona recogida de forma objetiva y estandarizada. Los psicólogos y otros profesionales recogen esas muestras de conducta porque a partir de ellas pueden hacer inferencias fundadas acerca del comportamiento y funcionamiento cognitivo de las personas evaluadas. La primera condición para que un test sirva de base para llevar a cabo inferencias de interés es que la muestra de conducta recogida sea precisa, es decir, que los errores cometidos en la medición sean aceptables, pues ninguna medición científica está exenta totalmente de error. Como se ha visto en los apartados precedentes, la tecnología psicométrica desarrollada para evaluar el grado de precisión de las mediciones realizadas con los test se denomina «fiabilidad».

La tecnología psicométrica encargada de mostrar que las inferencias hechas acerca del funcionamiento de las personas a partir de test son correctas es lo que denominamos «validez». Esta distinción entre fiabilidad y validez es razonable y útil, y se acepta sin mayor problema entre los especialistas. Ahora bien, tampoco sería descabellado considerar la fiabilidad como una primera fase del proceso de validación de un test, pues es difícil de imaginar que se puedan extraer inferencias enjundiosas a partir de test poco precisos. La cuestión central que nos ocupa en este apartado es clara: ¿de qué modo se comprueba que las inferencias hechas a partir de un test son correctas? En otras palabras, ¿cómo se procede para llevar a cabo el proceso de validación de las inferencias hechas a partir de las puntuaciones de un test?

Tiene que quedar muy claro desde el principio que, aunque se hable con frecuencia de *validar un test,* en sentido estricto no es el test lo que se valida, sino las inferencias que se hacen a partir de sus puntuaciones sobre determinados aspectos de la conducta de las personas. Por tanto, el resultado final de un proceso de validación no es llegar a decir de forma simplista que tal test es válido; las que son o no válidas son las inferencias hechas a partir del test con un determinado fin. Esto es natural, pues a partir de un test pueden hacerse inferencias de muy diverso tipo, de las cuales unas serán válidas y otras no; el proceso de validación consistirá precisamente en aportar datos y argumentos (evidencias) que permitan saber cuáles de las inferencias están fundadas, cuáles son válidas. ¿Cómo se aportan esos datos validantes? Es decir, ¿cómo se allegan la evidencia empírica y teórica necesarias para poder afirmar que determinadas inferencias realizadas son válidas? Responder a estos interrogantes es lo que constituye el meollo de la validez. Las respuestas, como no podía ser de otro modo, han ido variando a lo largo de la historia de la psicometría. Esta evolución queda muy bien reflejada en la literatura especializada, sobre todo en las sucesivas ediciones del manual clásico sobre medición psicológica y educativa editado sucesivamente por Lindquist (1951), Thorndike (1971), Linn (1989) y Brennan (2006), y, fundamentalmente, en las sucesivas ediciones de los estándares sobre los test publicados por la AERA, APA y NCME en 1954, 1966, 1974, 1985, 1999 y 2014, los cuales, en cierto modo, representan el consenso psicométrico oficial de cada época.

El concepto de validez, y por ende las prácticas de validación, han ido evolucionando desde unos

inicios marcadamente empíricos y operacionales a la situación actual, en la que se entiende la validez de una forma más amplia y comprensiva. Así, cuando Gulliksen (1950) sintetiza en su excelente manual lo esencial de la teoría clásica de los test de entonces, el problema de la validez se reduce a la correlación entre el test y el criterio a predecir. De modo que la tecnología psicométrica de la validez se centraba en el estudio de las correlaciones entre el test y los criterios a predecir, y las variables que modulaban esta relación, tales como la variabilidad de la muestra utilizada, la longitud del test, la fiabilidad del test y del criterio, o determinadas covariables. Nada que objetar; esta tecnología clásica sigue vigente en la actualidad. Lo que ocurre es que además de los datos relativos a la correlación test-criterio, el concepto de validez se ha ido ampliando paulatinamente. El trabajo pionero de Cronbach y Meehl (1955) sobre la validez de constructo alerta a teóricos, constructores y usuarios acerca de la importancia de ocuparse de la rigurosidad y entidad del constructo medido, además, obviamente, de trabajar con las correlaciones test-criterio. A partir de entonces, durante muchos años las vías esenciales para recoger datos en el proceso de validación de los test fueron el análisis de los contenidos de la prueba, las correlaciones test-criterio y la entidad de los constructos, lo que dio lugar a que se hablase de la santísisma trinidad de la validez: validez de contenido, validez de criterio y validez de constructo. Los estándares de 1985 ya dejan bien claro que, si bien esas tres vías de recogida de datos son legítimas, la validez es solo una y no hay razón alguna para que no se obtengan datos por cualquier otro camino complementario. En las propias palabras de los estándares de 1985 en su página 9: «Tradicionalmente las distintas formas de acumular evidencias sobre la validez se han agrupado en categorías denominadas validez de contenido, validez de criterio y validez de constructo. Estas categorías son útiles, como lo son otras categorizaciones más sofisticadas (por ejemplo dividir la validez de criterio en concurrente y predictiva), *pero el uso de estas categorías no quiere decir que haya distintos tipos de validez,* o que una estrategia determinada de validación sea mejor para cada inferencia específica o uso del test. No son posibles distinciones rigurosas entre cada categoría. Por ejemplo, datos relativos a la validez de contenido o a la validez de criterio son también de sumo interés para la validez de constructo. Una validación ideal incluye diferentes tipos de datos pertenecientes a las distintas categorías mencionadas». Ese es el planteamiento dominante sobre validez a partir de los años ochenta, que en el fondo no es otra cosa que subsumir los planteamientos sobre validez en el marco más general de la comprobación de hipótesis científicas. Validar las inferencias hechas a partir de las puntuaciones de los test es un caso particular de la validación de modelos e hipótesis científicas. En suma, el proceso de validación es unitario, y no hay tipos de validez; lo que hay son distintas vías y estrategias para aportar datos empíricos y teóricos (evidencias) que apoyen la pertinencia de las inferencias hechas a partir de las puntuaciones de las personas en los test. A continuación se comentan las estrategias más habituales para obtener evidencias empíricas en los procesos de validación.

1.1. Evidencias de validez

La psicometría ha conocido grandes avances en todas las ramas y la validez no es una excepción, si bien las novedades en este campo no han sido tan espectaculares como en otros. Como ya se ha señalado, se mantiene la filosofía general de la validez como un planteamiento unitario (Messick, 1980, 1988, 1989), aunque se utilicen distintas aproximaciones para obtener datos relevantes para la validación de las inferencias. Validar un test puede considerarse un caso particular de la comprobación de hipótesis científicas, pero no existe un método científico claro y universal (Weinberg, 2003) que aplicado de forma algorítmica dé solución a todos los problemas, lo cual tampoco quiere decir que todo vale. Este es un planteamiento correcto y teóricamente justificado, pero, como señala Brennan (1998, 2001), si bien la noción de una validez unitaria es muy sugerente teóricamente, hasta la fecha no ha mostrado una gran utilidad práctica de cara a los procesos reales de validación. Los constructores y usuarios de los test reclaman reglas más específicas que les permitan allegar datos que les ayuden a validar sus inferencias. Las tres vías clásicas para la recogida de datos, a saber, contenidos, relaciones con el criterio y constructo, siguen siendo feraces, por supuesto, pero algunas otras se han ido aña-

diendo en este proceso de construcción de la validez. Repasamos brevemente a continuación las más habituales, siguiendo aquellas expresamente citadas en los estándares (AERA, APA y NCME, 1999, 2014), pero dejando bien claro que cualesquiera otras son igualmente legítimas si se obtienen siguiendo los cánones habituales de la metodología científica, no hay ninguna razón para limitarse a las cinco que aquí se comentan.

Evidencias de contenido

Si los ítems que componen una prueba no representan adecuadamente el constructo que se pretende evaluar, difícilmente podrán ser correctas las inferencias que se hagan a partir del test. Todo proceso de validación ha de comenzar por la inexcusable tarea de comprobar la pertinencia de los contenidos; si esta falla, todo lo demás, por muy sofisticado técnicamente que sea, tiene los pies de barro. Algo tan elemental se olvida con cierta frecuencia, basando a veces la selección de ítems en criterios meramente estadísticos a posteriori. A la hora de llevar a cabo la validación de los contenidos han de comprobarse al menos dos aspectos vitales: la definición del constructo a evaluar y su correcta representación en el test. La definición ha de hacerse de forma operativa de modo que sea susceptible de someterse a prueba y sea posible derivar indicadores empíricos para su medición. No hay reglas universales para llevar a cabo una definición adecuada, depende en gran parte del constructo a medir. No es lo mismo, por ejemplo, definir las variables de tipo educativo o profesional, donde los dominios suelen estar bien acotados, que variables típicamente psicológicas como la extraversión o la inteligencia. Definido el constructo, la representación se refiere al grado en el que los ítems que componen el test representan todos los aspectos del constructo a medir.

Para lograr estos dos objetivos puede procederse de forma analítica y racional, mediante la utilización de expertos en la temática a evaluar, o bien usar técnicas estadísticas tras la aplicación de la prueba. Lo más recomendable es empezar con los expertos y complementar sus opiniones con los análisis estadísticos. A partir de los datos proporcionados por los expertos pueden obtenerse diversos indicadores cuantitativos de sus juicios; véanse, por ejemplo, Aiken (1980), Hambleton (1980, 1984), Popham (1992), Sireci y Geisinger (1992, 1995) y Deville (1996), entre otros. Para un tratamiento en profundidad de la problemática implicada en la validez de contenido y sus avatares históricos pueden consultarse los trabajos de Sireci (1998a y b, 2003), Kane (2006b) o Sireci y Faulkner-Bond (2014).

En suma, y como bien señala Sireci (1998a), sean cuales sean los debates teóricos sobre validez, que son muchos, y para seguir, en la práctica las evidencias de validez basadas en los contenidos son fundamentales, tal como lo recogen con justicia los últimos estándares (AERA, APA y NCME, 2014). Un aspecto importante de la validez de contenido es el que se refiere a la necesidad de que el test parezca, dé la impresión a las personas evaluadas, de que es adecuado y tiene sentido para medir lo que se pretende (Muñiz, 2003, 2004). Se trata de un tipo de evidencia sobre la validez de carácter menor, pero en determinadas circunstancias podría llevar a las personas evaluadas a desmotivarse para contestar la prueba si considerasen que aquello por las razones que sea no les parece serio. Tanto quien construye una prueba como quien ha de seleccionarla para su aplicación harían bien en asegurarse de que las tareas incluidas en la prueba, así como su apariencia, resultan aceptables para las personas evaluadas. Buenos análisis sobre la validez aparente pueden consultarse en Turner (1979), Friedman (1983) o Nevo (1985). En suma, hoy como ayer, y seguro que también mañana, todo proceso de validación comienza por la base, por los contenidos del test; después vendrá todo lo demás.

Procesos de respuesta

Las personas evaluadas mediante un test obtienen una determinada puntuación en los ítems y en el test y todas las inferencias que se hacen parten de esos datos. Cuanto más conozcamos acerca de los procesos que llevan a una persona a obtener una determinada puntuación, mejor comprenderemos el constructo medido y mayor control tendremos sobre las posibles predicciones. Los datos que se puedan aportar sobre estos procesos de respuesta constituyen una apoyatura excelente en el proceso de validación de la prueba; incluso podría afirmarse que en su ausencia no se puede hablar de una vali-

dación en profundidad. Nótese que estamos ante la tarea clásica propuesta por Cronbach (1957, 1975) de unir los esfuerzos de los enfoques diferencial y general para entender cabalmente la conducta humana. Las estrategias para aportar datos sobre los procesos subyacentes a las respuestas de las personas a los ítems de los test son muy variadas, si bien siempre se basan en el análisis de las respuestas individuales de las personas. Estas estrategias pueden ir desde preguntar a las propias personas acerca de su proceder y observar los pasos sucesivos (cuando es posible) que les conducen al resultado final hasta utilizar observadores expertos o analizar de forma experimental los procesos básicos y componentes implicados en la respuesta de cada ítem.

La emergencia del paradigma cognitivo en los años sesenta levantó grandes expectativas acerca de la posibilidad de poder dar cuenta de las respuestas de las personas a los ítems de los test de aptitudes, en especial de inteligencia. La literatura generada ha sido abundantísima, habiéndose estudiado exhaustivamente procesos tan diversos como el tiempo de reacción, la memoria, el tiempo de inspección o los potenciales evocados, solo por citar algunos. Tras cincuenta años de predominio del paradigma cognitivo en psicología, y pasados los primeros entusiasmos, hay que decir que no se ha avanzado mucho en el conocimiento de los procesos explicativos de las respuestas de las personas a los ítems. Seguimos sabiendo más acerca de las predicciones que se pueden hacer a partir de las puntuaciones en los test que sobre los procesos reales que hacen que unas personas resuelvan con soltura los ítems y otras lo hagan con dificultad. La esperada fecundación de la psicología diferencial por la psicología general sigue pendiente en gran medida. Analizar las causas profundas de este estado de cosas nos llevaría lejos, fuera del alcance de este libro, y es que la comprensión cabal de los procesos cognitivos que subyacen a la medición de las aptitudes y otras variables no está exenta de serios problemas (Prieto y Delgado, 1999).

En cualquier caso, la dificultad de aportar datos sobre los procesos implicados en la resolución de los ítems no debe disuadirnos de intentarlo, pues una validación en profundidad solo se conseguirá cuando se logren integrar coherentemente las puntuaciones obtenidas por las personas con los procesos seguidos para obtenerlas. Un buen análisis sobre estos aspectos de la validación puede verse en Padilla y Benítez (2014).

Estructura interna del test

Los datos sobre la estructura interna del test pretenden evaluar en qué medida el test constituye un constructo coherente y riguroso y no se trata simplemente de un conjunto espurio de ítems. Un test puede estar diseñado para constituir una o varias dimensiones, depende en cada caso de la definición operacional del constructo a medir. La evaluación de la dimensionalidad es uno de los tópicos con mayor tradición psicométrica, pues muchos de los modelos psicométricos más habituales en la práctica asumen que el constructo evaluado es unidimensional. La unidimensionalidad matemáticamente perfecta solo existe en la mente de quienes construyen y analizan los test; por tanto, trátase de ver en qué medida es aceptable la unidimensionalidad mostrada por los datos empíricos. En otras palabras, hay que asegurarse de la robustez de los modelos psicométricos utilizados a violaciones del supuesto de unidimensionalidad. Por ejemplo, diversos trabajos muestran que los modelos logísticos de teoría de respuesta a los ítems son bastante robustos a violaciones moderadas de la unidimensionalidad (Muñiz y Cuesta, 1993). Aunque autores como Hattie (1984, 1985) describen más de ochenta indicadores de unidimensionalidad, los más populares siguen siendo los derivados del análisis factorial, si bien otras muchas alternativas son actualmente posibles, tales como el uso de los modelos de ecuaciones estructurales (Gómez, 1996; Muthén, 1988; Pitoniak, Sireci y Luecht, 2002). Véanse buenas revisiones sobre dimensionalidad en Cuesta (1996) y Elosua y López (2002). Tal vez convenga recordar, dada la popularidad del coeficiente alfa de Cronbach para evaluar la fiabilidad, que si bien este se basa en la consistencia interna de la prueba, no puede tomarse sin más como un indicador de la dimensionalidad. No puede hablarse tampoco de una dimensionalidad intrínseca e invariante de una prueba, ya que esta puede variar con el tipo de muestra, e incluso viene afectada por el formato de los ítems (García-Cueto, Muñiz y Lozano, 2002).

Dentro de este apartado relativo a la estructura interna pueden ubicarse los trabajos encaminados a evaluar el funcionamiento diferencial de los ítems

(DIF). Estos análisis tratan de asegurar que los ítems funcionan de forma similar para diferentes grupos, no favoreciendo o perjudicando a unos grupos frente a otros. Nótese que datos sobre este funcionamiento de los ítems son claves para poder apoyar la validez y universalidad de una prueba. Seguramente la tecnología para la evaluación del DIF ha sido uno de los capítulos de la psicometría que más atención ha recibido durante los últimos años, habiendo llegado a soluciones técnicas muy satisfactorias para la evaluación eficiente del DIF. Pueden consultarse buenas exposiciones y análisis en Holland y Wainer (1993), Camilli y Shepard (1994), Fidalgo (1996), Fidalgo y Muñiz (2002) o Hidalgo y López-Pina (2000). Una asignatura pendiente de esta tecnología es la detección del DIF cuando la muestra utilizada es poco numerosa, en cuyo caso las técnicas convencionales, tales como el método Mantel-Haenszel, no funcionan todo lo bien que sería de desear (Muñiz, Hambleton y Xing, 2001). Un análisis sobre las evidencias basadas en la estructura interna de las pruebas puede verse en Ríos y Wells (2014).

Las estrategias para la obtención de datos comentadas hasta ahora se centraban en aspectos internos del test, bien fuese su contenido, los procesos implicados en las respuestas a los ítems o la estructura interna del test. A partir de ahora se comentan nuevas estrategias de obtención de evidencias, relativas a la conexión del test con distintas variables externas a él.

Relaciones con otras variables

a) *Converger y discriminar*

Un test diseñado para medir un determinado constructo no suele estar solo en el mundo; ese mismo constructo puede ser evaluado por muy diversos procedimientos más o menos similares a nuestro test. Si el constructo es sólido, tiene entidad y no es meramente espurio, las distintas mediciones que se hagan de él por el procedimiento que sea han de ser similares, han de converger, han de estar correlacionadas, han de mostrar, en suma, validez convergente; nada más natural. Análogamente, si distintos constructos se evalúan utilizando procedimientos parejos, no hay razón para esperar que dichas mediciones converjan; deberían divergir, discriminar un constructo del otro. En psicología no siempre ha ocurrido esto con todos los constructos utilizados por los psicólogos, como bien dieron cuenta de ello Campbell y Fiske (1959) en su trabajo pionero, formulando una tecnología, la matriz multirrasgo-multimétodo, para someter a prueba la existencia de evidencias de validez convergente y discriminante. Desde el trabajo de Campbell y Fiske se han seguido numerosas propuestas para analizar estadísticamente los datos provenientes de dichas matrices, buenos tratamientos de los cuales pueden consultarse, por ejemplo, en Browne (1984), Marsh (1988), Schmitt y Stults (1986), Kenny (1994) o Hernández y González-Romá (2000).

Aportar datos sobre el grado en el que un test converge con otras mediciones del mismo constructo, o diverge con aquellas de constructos diferentes, sigue siendo fundamental en su proceso de validación. En la práctica el problema radica en la dificultad y carestía en tiempo y dinero que suele implicar la obtención de los datos necesarios para llevar a cabo este tipo de análisis. Las evidencias de validez convergente y discriminante pueden obtenerse a partir de los datos proporcionados por la así llamada matriz multirrasgo-multimétodo, que no es otra cosa que lo que indica su nombre, a saber, una matriz de correlaciones en la que aparecen varios rasgos psicológicos (constructos) medidos por varios métodos. Dícese haber validez convergente si las correlaciones entre las medidas de un rasgo por distintos métodos son elevadas, es decir, las medidas de un mismo rasgo convergen, aunque se hayan hecho por diferente método. La validez discriminante se refiere a que las correlaciones anteriores entre las medidas del mismo rasgo por distintos métodos han de ser claramente superiores a las correlaciones entre las medidas de distintos rasgos por el mismo método. La idea de Campbell y Fiske (1959), aunque no era nueva en el ámbito de la psicología, sistematizada de este modo adquirió rápidamente gran difusión y popularidad, pues ya estaba latente en amplios sectores de la psicología la necesidad de garantizar que las teorías y constructos psicológicos al uso no eran meros artefactos emanados de un determinado método de medida que se desvanecían al variar este, como así comprobaron en numerosos casos Campbell y Fiske (1959) al revisar la literatura.

A modo de ejemplo, se presenta a continuación una matriz multirrasgo-multimétodo (tabla 3.1) en la

TABLA 3.1

		Extraversión			Liderazgo			Inteligencia social		
		AI	OS	EP	AI	OS	EP	AI	OS	EP
Extraversión	AI	0,80								
	OS	0,70	0,80							
	EP	0,60	0,70	0,90						
Liderazgo	AI	0,20	0,00	0,10	0,80					
	OS	0,00	0,00	0,00	0,70	0,80				
	EP	0,10	0,00	0,10	0,60	0,60	0,80			
Inteligencia social	AI	0,20	0,00	0,00	0,10	0,10	0,30	0,90		
	OS	0,00	0,10	0,10	0,10	0,20	0,10	0,70	0,80	
	EP	0,00	0,10	0,20	0,30	0,10	0,20	0,70	0,60	0,80

que tres rasgos (extraversión, liderazgo e inteligencia social) se midieron cada uno por tres métodos (autoinforme, observación sistemática y encuesta a profesores) en una muestra de escolares de 12 años.

En la diagonal principal aparecen las correlaciones de los test consigo mismos, esto es, los coeficientes de fiabilidad, todos ellos iguales o mayores que 0,80. La validez convergente, valores adyacentes a la diagonal principal, también es aceptable, con valores iguales o superiores a 0,60 en todos los casos. Asimismo, existe una clara validez discriminante, pues la máxima correlación entre medidas de distinto rasgo por el mismo método es 0,20. Naturalmente, la realidad suele presentarse no tan diáfana como este ejemplo ilustrativo. Para un estudio exhaustivo de los modelos de análisis de la matriz, véanse, por ejemplo, Browne (1984), Marsh (1988), Schmitt y Stults (1986) o Widaman (1985).

b) *La predicción del criterio*

Seguramente los test constituyen la tecnología más importante de la que disponen los psicólogos para ejercer su profesión e investigar numerosos aspectos de la conducta humana. Los test son tan utilizados porque, entre otras cosas, permiten hacer predicciones precisas sobre aspectos clave del funcionamiento humano. Pues bien, a la base de esas predicciones están las correlaciones entre el test y la variable a predecir, el criterio. La correlación entre el test y el criterio se denomina «coeficiente de validez» y es el dato que históricamente se ha propuesto en primer lugar para la validación de los test. Después vino todo lo demás, la validez de contenido (Cureton, 1951), la de constructo (Cronbach y Meehl, 1955) y las propuestas unificadoras de la validez que predominan en nuestros días (Messick, 1989). Todo lo relativo a los coeficientes de validez está bien tratado en los manuales clásicos como el de Gulliksen (1950), con aportaciones posteriores importantes derivadas del uso generalizado de las técnicas multivariadas, tales como la regresión múltiple, el análisis factorial o el análisis discriminante, entre otras. Véase una síntesis en Dunbar y Ordman (2003). Seguramente el problema más insidioso en este contexto es el de la evaluación precisa del propio criterio (Yela, 1990). En algunos casos su estimación no ofrece dificultad mayor, como ocurre con los test referidos al criterio (tanto educativos como profesionales), donde el domino viene acotado de forma precisa y operativa. En esos casos obtener la correlación test-criterio no conlleva mayores dificultades. Ahora bien, en el caso de algunas variables psicológicas, dar con un criterio adecuado, y medirlo con precisión, tórnase tarea poco menos que imposible. Piénsese, por ejemplo, en los criterios para la validación de los test de inteligencia: casi ninguno de ellos está exento de polémica. Algunas recomendaciones de interés para la medición del criterio pueden verse en Thorndike (1982) o Crocker y Algina (1986). Erróneamente suele dedicarse mucha más atención a garantizar las propiedades psicométricas del test, descuidándose las del criterio, cuando en realidad ambos son, como mínimo, igual de relevantes a la hora de calcular los

coeficientes de validez. Con frecuencia las personas implicadas en los procesos de validación asumen como aproblemática la medición del criterio, lo cual está lejos de la realidad; piénsese, por ejemplo, en criterios habituales tales como los juicios de expertos, supervisores o profesores, los cuales vienen afectados por numerosas fuentes de error que es necesario estimar.

La distinción clásica de validez concurrente, predictiva o retrospectiva, en función de que la medición del criterio se haga a la vez que el test, posterior o previamente, sigue siendo práctica para organizar los datos; además cada situación conlleva estrategias diferentes de medición del criterio. La visión unificadora de la validez que predomina actualmente desde el punto de vista conceptual no debe confundirnos y distraer nuestra atención sobre las correlaciones test-criterio; si estas no se aportan en el proceso de validación, poca o ninguna rentabilidad aplicada se le va a sacar a la prueba. El sino del test que no logra predecir un criterio de interés corre parejo al del buey que no ara.

c) *Generalización de la validez*

La pretensión de que los resultados hallados en cualquier ámbito científico sean universales, es decir, se puedan generalizar en condiciones diferentes a las que fueron hallados, constituye una premisa científica básica, y los datos obtenidos en los procesos de validación no son una excepción. La generalización hay que probarla, no se puede dar por supuesta, de modo que han de obtenerse datos y aportar argumentos para estar seguros de que las correlaciones test-criterio obtenidas en determinadas condiciones se mantienen en otras condiciones no estrictamente iguales, es decir, son generalizables. La variación de situaciones es prácticamente ilimitada, de modo que el aporte de datos que avalen la generalización constituye un proceso de acumulación progresiva. No obstante, no todas las variaciones circunstanciales tienen la misma entidad, y es tarea del usuario de los test y del constructor explicitar aquellos más relevantes para cada caso. Así, por ejemplo, en los estándares de la AERA, APA y NCME (2014) se mencionan cinco situaciones que pueden incidir en la generalización de los coeficientes de validez: diferencias en la forma en la que se mide el constructo predictor, el tipo de trabajo o currículum implicados, el tipo de medida del criterio utilizada, el tipo de personas evaluadas y el momento temporal en el que se lleva a cabo el estudio. Todos esos parámetros y otros muchos pueden variar de unos casos a otros, por lo que hay que ir acumulando datos empíricos acerca de la pertinencia de las generalizaciones. El metaanálisis se ha ido imponiendo como forma estándar de análisis de los datos proporcionados por las investigaciones; ahora bien, sus resultados para un caso particular no deben tomarse de forma ingenua, y es necesario asegurarse de que los trabajos incluidos en el metaanálisis son equiparables a la situación que nos ocupa en cada momento. Véase una buena revisión crítica en Murphy (2003).

Aparte de los cinco aspectos mencionados, un factor muy estudiado ha sido la incidencia del entrenamiento para hacer los test *(coaching)* en sus propiedades psicométricas (Allalouf y Shakhar, 1998; Anastasi, 1981; Jones, 1986; Linn, 1990; Martínez-Cardeñoso, García-Cueto y Muñiz, 2000; Messick y Jungeblut, 1981; Powers, 1985, 1986, 1993). Los resultados parecen indicar con claridad que los entrenamientos sistemáticos tienden a mejorar en cierto grado las puntuaciones de las personas en los test entrenados, eso sí, con importantes fluctuaciones en función del tipo de programa de entrenamiento, las horas invertidas y, sobre todo, el tipo de test. Sin embargo, no se dispone de datos concluyentes sobre la incidencia del entrenamiento en la fiabilidad y validez de los test.

Consecuencias del uso de los test

La última estrategia de recogida de datos en el proceso de validación propuesta por los recientes estándares (AERA, APA y NCME, 2014) es la inclusión de las consecuencias del uso de los test en el proceso de validación. Esta propuesta fue incluida por primera vez en los estándares de 1999. El debate sobre lo que ha dado en llamarse «validez consecuencial» se aviva a raíz del influyente trabajo de Messick (1989) en la tercera edición del libro *Educational Measurement,* donde propone ampliar el marco conceptual de la validez para dar cabida en él a las consecuencias del uso de los test. Nadie había dudado nunca, que se sepa, de la gran importancia que tiene ocuparse del uso adecuado de los test y de las consecuencias de su utilización, pero de

ahí a incluir estos aspectos dentro del marco científico de la validez había un trecho, que Messick propone caminar. Su propuesta cala en la comunidad psicométrica dominante hasta el punto de ser incluida en los estándares de 1999. Bien es verdad que no hay unanimidad al respecto, siendo recomendables los trabajos de Shepard (1997) y Linn (1997) a favor, y los de Popham (1997) y Mehrens (1997) en contra. La literatura generada es abundante; véase por ejemplo el monográfico de la revista *Educational Measurement: Issues and Practice* (Green, 1998; Lane, Parke y Stone, 1998; Linn, 1998; Moss, 1998; Reckase, 1998; Taleporos, 1998).

El meollo del debate se centra fundamentalmente en si es apropiado o no incluir las consecuencias sociales del uso de los test en el marco de la validez. De lo que nadie duda es de la importancia de estas y de la necesidad de ocuparse de ellas por parte de los distintos agentes implicados en la utilización de los test, tales como autores, constructores, distribuidores, usuarios, personas evaluadas e instituciones contratantes (Haertel, 2002; Kane, 2002; Lane y Stone, 2002; Ryan, 2002). Al incluir las consecuencias sociales en el marco de la validez, se corre el riesgo de introducir por la puerta de atrás los planteamientos sociales y políticos en el estudio de la validez, que debería reservarse para los argumentos científicos. Autores como Maguire, Hattie y Brian (1994) consideran que esta insistencia en incluir las consecuencias dentro del marco de la validez viene motivada en gran parte por las continuas refriegas legales que rodean a los test en Estados Unidos. Consideran que si bien esta postura pudiera reducir las batallas legales, también puede distraer a los constructores de su misión central, que no es otra que aportar datos de cómo el test representa al constructo medido.

Este debate puede resultar ajeno a muchos psicólogos, especialmente a los clínicos, y ello no es de extrañar, pues surge fundamentalmente en los programas americanos de test a gran escala en el ámbito de la medición educativa y algo menos en la orientación y selección de personal. En España estamos poco familiarizados con estos problemas, pues escasean los programas sistemáticos de aplicación de pruebas a nivel regional o nacional. Pero imagínese por un momento que a determinada edad escolar todos los niños fuesen evaluados por una prueba educativa, y que esa prueba tuviese repercusiones importantes para la vida académica de los estudiantes. En esta situación todas las partes implicadas, constructores de pruebas, estudiantes, padres, colegios y gobiernos, mirarían con lupa todo el proceso y sus consecuencias. Aparte del propio test, las consecuencias de su aplicación, positivas y negativas, también serían escrutadas. Por ejemplo, una consecuencia positiva sería que los colegios se verían presionados para que sus estudiantes mejorasen y puntuasen alto en la prueba. Una posible consecuencia negativa sería que los programas se ajustarían y centrarían en aquellos aspectos incluidos en la prueba, restringiendo así los objetivos de la enseñanza; es decir, se enseñaría para la prueba. No cabe duda de que se trata de dos consecuencias de interés que han de tenerse en cuenta; la cuestión que se debate es si han de ser incluidas en el marco de los estudios de validez o no.

Nótese que esta cuestión de la validez consecuencial no se identifica estrictamente con el uso inadecuado de los test, que sencillamente ha de evitarse, para lo cual las organizaciones nacionales e internacionales llevan a cabo muy diversas iniciativas; véanse, por ejemplo, Bartram (1998), Fremer (1996), Evers (1996), Evers et al. (2017), Muñiz (1997, 1998), Muñiz y Fernández-Hermida (2000), Muñiz, Prieto, Almeida y Bartram (1999) y Simner (1996). Una alternativa razonable sería incluir en esta tradición del uso adecuado de los test todo lo relativo a las consecuencias, pero hay quien considera que esto sería rebajar la importancia atribuida a las consecuencias, ya que incluidas en el capítulo de la validez, tienen garantizada una mayor cuota de pantalla en el debate psicométrico. Véase un análisis detallado del papel de las consecuencias en el proceso de validación en Padilla et al. (2007) o Lane (2014).

Comentarios finales

Para orientar en la práctica la obtención de evidencias empíricas relativas a las cinco vías descritas, contenidos, procesos, estructura interna, relaciones con otros test y consecuencias, los estándares técnicos (AERA, APA, NCME, 2014) proponen veinticinco directrices de gran interés, remitiendo a ellas a los lectores interesados. Además, tratamientos detallados sobre el proceso de validación pueden verse en Paz (1996), Elosua (2003), Kane (2006b, 2016), Lissitz (2009) o Zumbo y Chan (2014), y para un planteamiento más crítico, Markus y Borsboom

(2013). Tal como se ha expuesto, los planteamientos actuales sobre el proceso de validación de los test hacen especial hincapié en la necesidad de aportar datos empíricos y fundamentación teórica para justificar cualquier inferencia que se pretenda hacer a partir de las puntuaciones de los test. Si bien se ha ido refinando y sofisticando históricamente el concepto de validez, evolucionando hacia un planteamiento unitario, dentro del marco general de la metodología científica para la comprobación de hipótesis, el tipo de datos empíricos obtenidos en el proceso de validación ha permanecido más estable. No obstante, ello no ha sido en balde, pues el nuevo marco unitario permite interpretarlos de forma más integradora y significativa. La validación pasa así a ser conceptualizada como un caso particular de la metodología científica. La recogida de datos para someter a prueba las inferencias hechas a partir de los test conlleva la misma problemática que la comprobación de cualquier otra hipótesis científica. Las estrategias de recogida de datos clásicas son lícitas, por supuesto, pero no necesariamente exclusivas.

2. VALIDEZ Y FIABILIDAD

El coeficiente de validez es la correlación entre el test y el criterio, y constituye el dato esencial a la hora de obtener evidencias de validez del test relativas a las relaciones con otras variables. Para calcularlo hay que medir tanto el test como el criterio, por lo que la fiabilidad de estas mediciones influirá en el tamaño de la correlación obtenida entre ambos. En este apartado se analiza cómo influye la fiabilidad del test y del criterio en el coeficiente de validez. Veremos que, si se mejora la fiabilidad del test y del criterio, el coeficiente de validez aumentará; en qué grado lo hace es lo que se muestra en este apartado.

2.1. Fórmulas de atenuación

a) *Estimación del coeficiente de validez en el supuesto de que el test y el criterio tuviesen una fiabilidad perfecta*

Supóngase, por ejemplo, que un psicólogo ha calculado el coeficiente de validez de un test, y que dispone también de la fiabilidad tanto del test como del criterio. Una pregunta muy pertinente que se le plantea a continuación podría ser la siguiente: ¿cuál sería la validez del test en el supuesto de que tanto el test como el criterio tuviesen una fiabilidad perfecta? Es decir, en el supuesto de que careciesen de errores de medida.

La respuesta viene dada por la fórmula de atenuación (Spearman, 1904), denominación que hace referencia al hecho de que la validez empírica viene atenuada, reducida, disminuida, por la existencia de los errores de medida, existencia cuya fórmula permite corregir, o, más exactamente, permite hacer una estimación, según los supuestos del modelo, de cuál sería la validez si test y criterio careciesen de errores de medida. En ese caso de ausencia de errores, la validez del test vendría dada por la correlación entre las puntuaciones verdaderas de las personas en el test y sus verdaderas en el criterio:

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} \qquad [3.1]$$

donde:

ρ_{xy}: Coeficiente de validez empírico.
$\rho_{xx'}$: Coeficiente de fiabilidad empírico del test.
$\rho_{yy'}$: Coeficiente de fiabilidad empírico del criterio.

Efectivamente, por definición:

$$\rho_{v_x v_y} = \frac{\text{cov}\,(V_x, V_y)}{\sigma_{v_x}\sigma_{v_y}}$$

Ahora bien, según [1.6], la covarianza entre las puntuaciones verdaderas es igual a la covarianza entre las empíricas:

$$\text{cov}\,(V_x, V_y) = \text{cov}\,(X, Y) = \rho_{xy}\sigma_x\sigma_y$$

luego:

$$\rho_{v_x v_y} = \frac{\sigma_{xy}}{\sigma_{v_x}\sigma_{v_y}} = \frac{\rho_{xy}\sigma_x\sigma_y}{\sigma_{v_x}\sigma_{v_y}} = \frac{\rho_{xy}}{\left(\frac{\sigma_{v_x}}{\sigma_x}\right)\left(\frac{\sigma_{v_y}}{\sigma_y}\right)} =$$

$$= \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

que es la fórmula propuesta.

Ejemplo

El coeficiente de validez de una escala de motivación de logro resultó ser 0,60; su coeficiente de fiabilidad, 0,64, y el coeficiente de fiabilidad del criterio, 0,81. ¿Cuál se estima que sería el coeficiente de validez de la escala en el supuesto de que tanto la escala como el criterio careciesen de errores de medida?

$$\rho_{v_x v_y} = \frac{0,60}{\sqrt{0,81}\sqrt{0,64}} = 0,83$$

El incremento del coeficiente de validez sería considerable, pasando de 0,60 a 0,83, en el caso de que la escala y el criterio tuviesen una fiabilidad perfecta.

b) *Estimación del coeficiente de validez en el caso de que el test tuviese una fiabilidad perfecta*

La estimación anterior puede realizarse también en el caso de que solo el test carezca de errores de medida, en cuyo caso la validez vendría dada por:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}} \qquad [3.2]$$

puesto que:

$$\rho_{v_x y} = \frac{\mathrm{cov}(V_x, Y)}{\sigma_{v_x}\sigma_y} = \frac{\mathrm{cov}(X, Y)}{\sigma_{v_x}\sigma_y} = \frac{\rho_{xy}\sigma_x\sigma_y}{\sigma_{v_x}\sigma_y} =$$

$$= \frac{\rho_{xy}}{\frac{\sigma_{v_x}}{\sigma_x}} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}}$$

En el ejemplo anterior, si solo el test careciese de errores de medida, la estimación del coeficiente de validez sería:

$$\rho_{v_x y} = \frac{0,60}{\sqrt{0,64}} = 0,75$$

Lógicamente, el valor que toma ahora el coeficiente de validez (0,75) es menor que el estimado en el apartado anterior (0,83), puesto que allí estaban libres de error tanto test como criterio, mientras que aquí solo lo está el test.

c) *Estimación del coeficiente de validez en el caso de que el criterio tuviese una fiabilidad perfecta*

Análogamente a los casos anteriores, cuando es el criterio el que únicamente carece de errores de medida, la estimación del coeficiente de validez viene dada por:

$$\rho_{x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{yy'}}} \qquad [3.3]$$

puesto que:

$$\rho_{x v_y} = \frac{\mathrm{cov}(X, V_y)}{\sigma_x\sigma_{v_y}} = \frac{\mathrm{cov}(X, Y)}{\sigma_x\sigma_{v_y}} = \frac{\rho_{xy}\sigma_x\sigma_y}{\sigma_x\sigma_{v_y}} =$$

$$= \frac{\rho_{xy}}{\frac{\sigma_{v_y}}{\sigma_y}} = \frac{\rho_{xy}}{\sqrt{\rho_{yy'}}}$$

Para los datos del ejemplo anterior:

$$\rho_{x v_y} = \frac{0,60}{\sqrt{0,81}} = 0,67$$

d) *Generalización de las fórmulas de atenuación*

A partir de las fórmulas de atenuación anteriores, puede hacerse una generalización para el supuesto en el que test y/o criterio, aun sin carecer totalmente de errores de medida, se les rebajen en cierto grado. Es decir, puede tener interés estimar cuál sería el coeficiente de validez de un test en el supuesto de que se lograsen ciertas mejorías en su fiabilidad, en la del criterio o en ambas. La fórmula que nos da tal estimación es la siguiente:

$$\rho_{xy} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} \qquad [3.4]$$

donde las letras mayúsculas se refieren a las fiabilidades mejoradas.

Efectivamente, según [3.1]:

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} \qquad (1)$$

Ahora bien, si se lograse mejorar en cierto grado la fiabilidad del test y del criterio, es decir, si se lograse reducir sus errores de medida, la correlación entre las puntuaciones verdaderas del test y las del criterio ($\rho_{v_x v_y}$) seguiría siendo la misma, puesto que lo que se modifica son los errores, que según el modelo no covarían con las verdaderas; luego:

$$\rho_{v_x v_y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}} \qquad (2)$$

Igualando (1) y (2):

$$\frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}$$

Despejando ρ_{XY}

$$\rho_{XY} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

que es la fórmula propuesta en [3.4], y que permite estimar la validez de un test cuando se han introducido mejoras en su fiabilidad y en la del criterio.

EJEMPLO

El coeficiente de validez de un test es 0,60; su fiabilidad, 0,70, y la fiabilidad del criterio, 0,80. ¿Cuál sería el coeficiente de validez del test en el supuesto de que se lograse elevar la fiabilidad del test a 0,75 y la del criterio a 0,90?

$$\rho_{XY} = \frac{0{,}60\sqrt{0{,}75}\sqrt{0{,}90}}{\sqrt{0{,}70}\sqrt{0{,}80}} = 0{,}66$$

El coeficiente de validez pasaría de 0,60 a 0,66.

Casos particulares de [3.4]

Cuando únicamente se mejora la fiabilidad del criterio, o la del test, pero no ambas, la expresión dada en [3.4] se puede simplificar.

— Mejora en la fiabilidad del criterio:

Puesto que en este caso la fiabilidad del test no se altera, $\rho_{xx'} = \rho_{XX'}$, anulándose ambos términos en el numerador y denominador:

$$\rho_{xY} = \frac{\rho_{xy}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}} \qquad [3.5]$$

— Mejora en la fiabilidad del test:

En este segundo caso, lo que permanece constante es la fiabilidad del criterio, $\rho_{yy'} = \rho_{YY'}$, que al anularse en el numerador y denominador reduce [3.4] a:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}} \qquad [3.6]$$

2.2. Valor máximo del coeficiente de validez

A partir de [3.1] es inmediato que el coeficiente de validez de un test es menor o igual que su índice de fiabilidad:

$$\rho_{xy} \leqslant \rho_{xv} \qquad [3.7]$$

Efectivamente, según [3.1]:

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

Ahora bien, dado que $\rho_{v_x v_y} \leqslant 1$

$$\frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} \leqslant 1$$

luego,

$$\rho_{xy} \leqslant \sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}$$

y, en consecuencia,

$$\rho_{xy} \leqslant \sqrt{\rho_{xx'}}$$

ya que el valor máximo de $\rho_{yy'}$ es 1. Ahora bien,

$$\sqrt{\rho_{xx'}} = \rho_{xv}$$

luego:

$$\rho_{xy} \leqslant \rho_{xv}$$

Esta expresión también nos indica que el coeficiente de validez puede ser mayor que el coeficiente de fiabilidad, ya que si

$$\rho_{xy} \leqslant \sqrt{\rho_{xx'}}$$

ρ_{xy} será igual a $\rho_{xx'}$ cuando este valga 1. En el resto de los casos ρ_{xy} podría ser mayor o igual que $\rho_{xx'}$, puesto que la raíz cuadrada de un número menor que 1 es mayor que dicho número.

Por ejemplo, si el coeficiente de fiabilidad de un test es $\rho_{xx'} = 0{,}81$:

$$\rho_{xy} \leqslant \sqrt{0{,}81}$$
$$\rho_{xy} \leqslant 0{,}90$$

Es decir, el valor máximo de ρ_{xy} es 0,90, mientras que $\rho_{xx'}$ es 0,81; luego ρ_{xy}, ciertamente, puede ser mayor que $\rho_{xx'}$.

2.3. Validez y longitud del test

La fiabilidad de un test puede mejorarse por varios caminos. Uno muy típico, como ya se ha visto en su momento, es aumentando su longitud. A su vez, la mejora de la fiabilidad del test repercute favorablemente sobre su validez, repercusión que viene dada precisamente por la fórmula [3.4]. Es claro, por tanto, que la longitud del test influye sobre su validez; al aumentar la longitud, mejora la validez. La relación exacta entre longitud y validez viene dada por la fórmula:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1 + (n-1)\rho_{xx'}}} \qquad [3.8]$$

donde n es el número de veces que se aumenta el test, ρ_{xy} el coeficiente de validez y $\rho_{xx'}$ el de fiabilidad.

Su obtención es inmediata, sustituyendo en [3.6] el valor de $\rho_{xx'}$ dado por la fórmula de Spearman-Brown expuesta en [2.20]. Según [2.20]:

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}}$$

Sustituyendo en [3.6]:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{\dfrac{n\rho_{xx'}}{[1 + (n-1)\rho_{xx'}]}}}{\sqrt{\rho_{xx'}}}$$

Simplificando $\sqrt{\rho_{xx'}}$ en el numerador y el denominador se obtiene directamente [3.8].

Bell y Lumsden (1980) aportan datos empíricos bastante favorables acerca del funcionamiento de esta fórmula.

Ejemplo

La fiabilidad de un test es de 0,80, y su validez, de 0,60. ¿Cuál sería el coeficiente de validez del test si se duplicase su longitud?

$$\rho_{Xy} = \frac{(0{,}60)\sqrt{2}}{\sqrt{1 + (2-1)(0{,}80)}} = 0{,}626$$

Despejando n de [3.8], se puede estimar cuánto habría que aumentar la longitud de un test para que alcanzase un coeficiente de validez determinado:

$$n = \frac{(1 - \rho_{xx'})\rho_{Xy}^2}{\rho_{xy}^2 - \rho_{Xy}^2\rho_{xx'}} \qquad [3.9]$$

EJEMPLO

La fiabilidad de un test que consta de 25 ítems es 0,70, y su validez, 0,80. ¿Cuánto habría que alargar el test para alcanzar una validez de 0,90?

$$n = \frac{(1 - 0{,}70)(0{,}90)^2}{(0{,}80)^2 - (0{,}90)^2(0{,}70)} = 3{,}33$$

Habría que alargar el test 3,33 veces, y dado que el test original constaba de 25 ítems, el nuevo alargado tendría que tener 84 ítems: [(25)(3,33) = 83,25]. Luego habría que añadir 59 ítems al original (84 − 25 = 59) para que alcanzase una validez de 0,90.

3. VALIDEZ Y VARIABILIDAD

El coeficiente de validez de un test se ha definido como la correlación entre el test y el criterio, y, como es bien sabido, la correlación entre dos variables tiende a aumentar con la variabilidad de la muestra; por tanto, a la hora de interpretar el coeficiente de validez, es de suma importancia conocer la variabilidad de la muestra en la que fue calculado. Como se verá más adelante, un mismo test puede tener un coeficiente de validez bien distinto dependiendo de la cuantía de la variabilidad (varianza) de la muestra en la que haya sido calculado.

En este apartado se analiza en qué grado viene afectado el coeficiente de validez por la variabilidad de la muestra. Para una mejor comprensión se considerarán tres casos según el número de variables implicadas:

— Dos variables.
— Tres variables.
— *n* variables (caso general).

3.1. Dos variables

Cuando se pretende calcular el coeficiente de validez, no es infrecuente que en psicología y educación se presente una situación como la que sigue. Supóngase un psicólogo al que se encomienda seleccionar unos pocos candidatos de entre un gran número de aspirantes a cierto trabajo. Amén de otras técnicas de selección de personal, seguramente utilizará alguna batería de test. Si pasado un tiempo tras la realización de la selección desease calcular la validez de la batería, no le quedaría otra opción que correlacionar las puntuaciones de los admitidos con

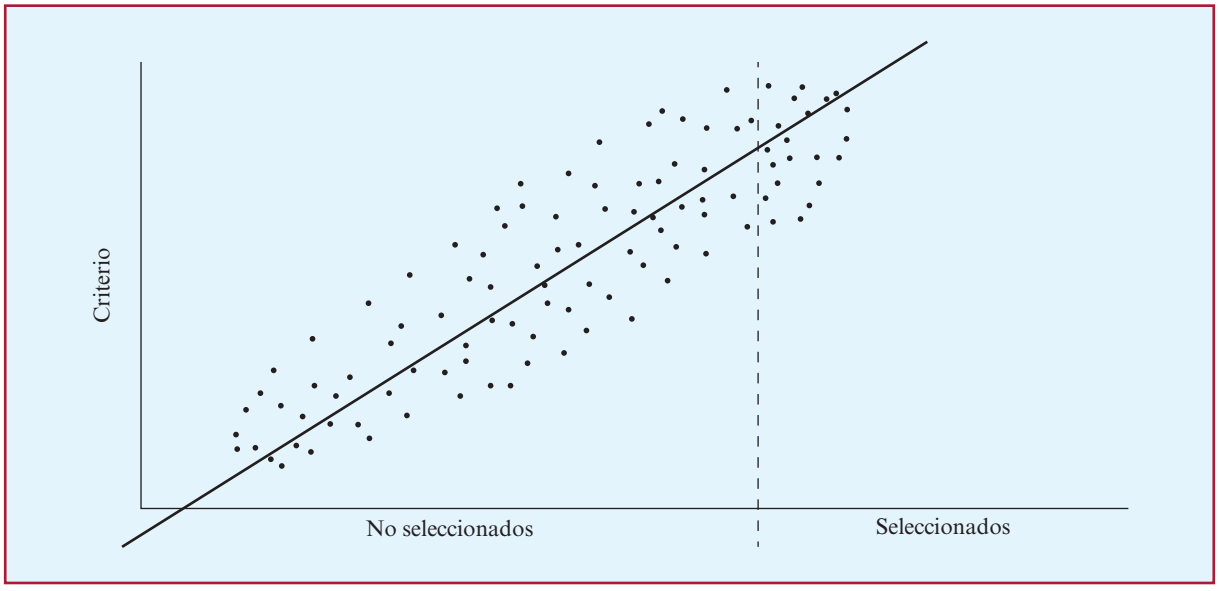


Figura 3.1.—Puntuaciones de una muestra de personas en una batería de test y en el criterio.

algún criterio de eficiencia en su trabajo. Nótese bien que el resultado así hallado no sería propiamente la validez de la batería, sería una estimación por defecto, puesto que ha sido calculada en el pequeño grupo de los admitidos y no en el total de los aspirantes. La variabilidad en el grupo de los seleccionados es mucho menor que en el grupo de aspirantes; luego la validez se verá notablemente rebajada. Véase en la figura 3.1 cómo el grupo de personas seleccionadas, que han superado la puntuación X en la batería, son mucho más homogéneas (menos variables) que el grupo total de aspirantes.

Ahora bien, ¿cómo saber cuál sería la validez de la batería si se hubiese calculado en el grupo total? Una solución, la mejor, sería admitir a trabajar a todos los aspirantes, puntuarles después de un tiempo en algún criterio de eficiencia y correlacionar estas puntuaciones con las de la batería. Naturalmente, este modo de proceder será casi siempre inviable en la práctica por razones obvias.

La alternativa será calcular el coeficiente de validez en el grupo de admitidos y, basándose en ciertos supuestos del modelo lineal clásico, hacer una estimación del valor que tomaría la validez si hubiese sido calculada en el grupo total de aspirantes. La precisión de la estimación dependerá de la pertinencia de los supuestos.

Supuestos

1. $$\frac{\rho_{xy}\sigma_y}{\sigma_x} = \frac{\rho_{XY}\sigma_Y}{\sigma_X}$$

2. $$\sigma_y\sqrt{1-\rho_{xy}^2} = \sigma_Y\sqrt{1-\rho_{XY}^2} \qquad [3.10]$$

Las letras mayúsculas se refieren al grupo de aspirantes, y las minúsculas, a los seleccionados.

Se asume:

1. Que la pendiente de la recta de regresión del criterio sobre el test es igual en ambos grupos.
2. Que el error típico de estimación también es igual en ambos grupos.

Bajo tales supuestos, de [3.10] se puede despejar ρ_{XY}, la validez en el grupo de aspirantes:

$$\rho_{XY} = \frac{\sigma_X\rho_{xy}}{\sqrt{\sigma_X^2\rho_{xy}^2 + \sigma_x^2 - \sigma_x^2\rho_{xy}^2}} \qquad [3.11]$$

Asimismo (despejando), se puede estimar cuál sería la variabilidad del criterio en el grupo total:

$$\sigma_Y = \sigma_y\sqrt{1 - \rho_{xy}^2 + \frac{\rho_{xy}^2\sigma_X^2}{\sigma_x^2}} \qquad [3.12]$$

Ejemplo

Se aplicó un test de rapidez perceptiva a un grupo de 2.000 aspirantes a pilotos de aviación. La desviación típica de los aspirantes en el test fue 20. Seleccionados 50 de los aspirantes, su desviación típica en el test fue 5, y la correlación entre las puntuaciones en el test y la eficacia como piloto tras un año de entrenamiento fue 0,25. ¿Cuál se estima que sería la validez del test si se hubiese calculado en el grupo de los aspirantes y no en el de los seleccionados, como así se hizo?

$$\rho_{XY} = \frac{(20)(0{,}25)}{\sqrt{(20)^2(0{,}25)^2 + 5^2 - 5^2(0{,}25)^2}} = 0{,}72$$

El cambio es bastante dramático: de 0,25 en el grupo de seleccionados se pasa a 0,72 en el grupo de aspirantes, y ello porque en el primero la desviación típica es 5, y en el segundo, 20.

En este contexto suele denominarse «variable directamente selectiva» a la utilizada explícitamente para seleccionar a las persoans. Lo más habitual en la práctica es que sea el test, pero también cabe la posibilidad de utilizar el criterio.

Por el contrario, variables indirectamente selectivas serían aquellas correlacionadas con la directamente selectiva. Si se lleva a cabo una selección mediante una determinada variable, la variabilidad en el grupo seleccionado disminuirá notablemente para esa variable, pero también lo hará para cualquier otra variable correlacionada con ella, es decir, se produce una selección indirecta incidental en las variables correlacionadas.

Según los datos que se asumen conocidos y desconocidos en los supuestos, se genera toda una ca-

suística que el lector puede encontrar expuesta con detalle en Gulliksen (1950).

Las fórmulas expuestas pueden usarse de forma general siempre que se reserve la X para la variable directamente selectiva, la Y para la indirectamente selectiva (independientemente de que sea el test o el criterio), las minúsculas para el grupo con los datos conocidos y las mayúsculas para el desconocido (independientemente de que el grupo sea el seleccionado o el total).

Sobre lo adecuado de los supuestos, no se anda sobrado de evidencia empírica confirmatoria, pero todo parece indicar que si la selección no es muy extrema, las fórmulas funcionan bien, e incluso infraestiman la validez en el grupo no seleccionado (Lord y Novick, 1968). En el caso de selección extrema, lo cual es bastante frecuente en la práctica, las fórmulas han de usarse con precaución, si bien Lee, Miller y Graham (1982) obtuvieron estimaciones ajustadas incluso para casos de selección extrema (10%).

3.2. Tres variables

Otra situación paradigmática en el ámbito de la psicología aplicada se produce cuando, una vez hecha la selección según lo expuesto en el apartado anterior, se plantea la posibilidad de que un nuevo test Z pueda reemplazar ventajosamente como selector al X que se viene utilizando. Una forma ingenua para tratar de averiguar si el nuevo test Z es preferible como selector al X podría consistir en aplicar el nuevo test Z a los seleccionados con el X, hallar los coeficientes de validez de ambos test (ρ_{xy}, ρ_{zy}) y compararlos. Sin embargo, esto no sería correcto, pues al comparar los valores de los coeficientes de validez obtenidos en el grupo de seleccionados se perjudicaría más al test con el que se hizo la selección, en el que la variabilidad será menor, que al nuevo. Lo apropiado será estimar el coeficiente de validez de ambos en el grupo de aspirantes (ρ_{XY}, ρ_{ZY}) y compararlos.

Para el cálculo de ρ_{XY} se utiliza [3.11], mientras que ρ_{ZY} viene dado por la expresión:

$$\rho_{ZY} = \frac{\rho_{zy} - \rho_{xz}\rho_{xy} + \dfrac{\rho_{xz}\rho_{xy}\sigma_X^2}{\sigma_x^2}}{\sqrt{\left(1 - \rho_{xz}^2 + \dfrac{\rho_{xz}^2\sigma_X^2}{\sigma_x^2}\right)\left(1 - \rho_{xy}^2 + \dfrac{\rho_{xy}^2\sigma_X^2}{\sigma_x^2}\right)}} \qquad [3.13]$$

donde, como en el caso de dos variables, X es el test directamente selectivo, Z el nuevo test e Y el criterio, reservándose las mayúsculas para el grupo total y las minúsculas para el de los seleccionados.

Ejemplo

Un cuestionario de habilidades sociales se utilizó para seleccionar 40 candidatos entre 1.000 aspirantes a encuestadores. La desviación típica de los aspirantes en el cuestionario fue 25, y la de los seleccionados, 6. Tras varios meses encuestando, la correlación entre las puntuaciones en el cuestionario y la eficiencia como encuestadores, medida según cierto criterio, fue de 0,30, mientras que la eficiencia encuestadora correlacionó 0,35 con un test de inteligencia general aplicado a los seleccionados. Por su parte, la correlación entre la inteligencia general y las habilidades sociales resultó ser de 0,60. A la luz de estos datos, ¿puede afirmarse que la inteligencia general predice la eficiencia encuestadora mejor que el cuestionario de habilidades sociales?

Según [3.11]:

$$\rho_{XY} = \frac{(25)(0{,}30)}{\sqrt{(25)^2(0{,}30)^2 + 6^2 - 6^2(0{,}30)^2}} = 0{,}80$$

Según [3.13]:

$$\rho_{ZY} = \frac{0{,}35 - (0{,}60)(0{,}30) + \dfrac{(0{,}60)(0{,}30)(25)^2}{6^2}}{\sqrt{\left[1 - (0{,}60)^2 + \dfrac{(0{,}60)^2(25)^2}{6^2}\right]}\sqrt{\left[1 - (0{,}30)^2 + \dfrac{(0{,}30)^2(25)^2}{6^2}\right]}} =$$

$$= 0{,}80$$

La respuesta es negativa; a pesar de que en el grupo seleccionado la correlación de la inteligencia general con el criterio era algo mayor, 0,35 frente a 0,30, en el grupo amplio resultaron iguales.

Los supuestos que subyacen para la obtención de [3.13] son análogos a los ya explicitados para el caso de dos variables, aquí generalizados a tres. Es

decir, se asume que las pendientes de las rectas de regresión de las variables indirectamente selectivas (Z, Y) sobre la variable directamente selectiva (X) son iguales en el grupo seleccionado y en el total (supuestos 1 y 2). Asimismo, se asume que son iguales sus errores típicos de estimación (supuestos 3 y 4), y que la correlación parcial entre el criterio (Y) y el nuevo test (Z), eliminando el efecto de X, es igual en ambos grupos (supuesto 5). Todo lo cual puede expresarse del siguiente modo:

1. $\dfrac{\rho_{xy}\sigma_y}{\sigma_x} = \dfrac{\rho_{XY}\sigma_Y}{\sigma_X}$
2. $\dfrac{\rho_{xz}\sigma_z}{\sigma_x} = \dfrac{\rho_{XZ}\sigma_Z}{\sigma_X}$
3. $\sigma_y\sqrt{1-\rho_{xy}^2} = \sigma_Y\sqrt{1-\rho_{XY}^2}$ [3.14]
4. $\sigma_z\sqrt{1-\rho_{xz}^2} = \sigma_Z\sqrt{1-\rho_{XZ}^2}$
5. $\dfrac{\rho_{zy}-\rho_{xz}\rho_{xy}}{\sqrt{(1-\rho_{xz}^2)(1-\rho_{xy}^2)}} = \dfrac{\rho_{ZY}-\rho_{XZ}\rho_{XY}}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{XY}^2)}}$

Trate el lector, a modo de ejercicio, de llegar a la expresión [3.13] despejando ρ_{ZY} a partir de los supuestos. En función de los datos conocidos-desconocidos de los supuestos, puede aparecer un desfile innumerable de casos posibles; véase, por ejemplo, Gulliksen (1950).

3.3. Caso general

Todo lo dicho para los casos de dos y tres variables puede generalizarse para el caso de n variables directamente selectivas y K indirectamente selectivas, en cuyo caso los supuestos expresados en forma matricial vendrían dados por:

1. $b_{yx} = b_{YX}$
2. $C_{ee} = C_{EE}$ [3.15]

El primer supuesto establece que los pesos de las variables directamente selectivas para predecir las indirectamente selectivas son iguales en ambos grupos, y el segundo, que las matrices de varianzas-covarianzas de los errores de estimación son iguales.

Aunque la casuística posible es variada, dependiendo de los datos que se consideren conocidos, en la práctica lo más usual será disponer de todos los datos en ambos grupos para las variables directamente selectivas y los de las variables indirectamente selectivas en el grupo seleccionado. Se ilustra este caso a continuación; véase Gulliksen (1950) para un tratamiento detallado.

Como se verá en el apartado siguiente, los pesos b vienen dados por:

$$\mathbf{b} = C_{xx}^{-1}C_{xy}$$

donde C_{xx}^{-1} es la inversa de la matriz de varianzas-covarianzas de las variables predictoras X (aquí directamente selectivas) y C_{xy} las covarianzas entre las directa e indirectamente selectivas.

Por otra parte:

$$C_{ee} = C_{yy} C'_{yx} C_{xx}^{-1} C_{xy}$$

Sustituyendo estos valores en los supuestos:

1. $C_{xx}^{-1}C_{xy} = C_{XX}^{-1}C_{XY}$
2. $C_{yy} - C'_{yx}C_{xx}^{-1}C_{xy} = C_{YY} - C'_{YX}C_{XX}^{-1}C_{XY}$

Del primero lo único que no se conoce es C_{XY}, que se puede despejar:

$$C_{XY} = C_{XX}C_{xx}^{-1}C_{xy} \qquad [3.16]$$

Sustituyendo en 2 el valor de C_{XY} y despejando C_{YY}:

$$C_{YY} = C_{yy} + C'_{yx}C_{xx}^{-1}C_{XX}C_{xx}^{-1}C_{xy} - C'_{yx}C_{xx}^{-1}C_{xy} \qquad [3.17]$$

Las expresiones [3.16] y [3.17] proporcionan, respectivamente, las covarianzas entre las variables directa e indirectamente selectivas, y las varianzas-covarianzas entre las indirectamente selectivas, ambas en el grupo amplio. Dado que la matriz de varianzas-covarianzas de las variables directamente selectivas (C_{XX}) se conoce, utilizando la fórmula de la correlación de Pearson puede estimarse cualquier coeficiente de validez en el grupo amplio [$\rho_{ZY} = \text{cov}(X, Y)/\sigma_X\sigma_Y$], los pesos b ($\mathbf{b} = C_{XX}^{-1}C_{XY}$), o la correlación múltiple.

EJEMPLO

De un total de 1.000 aspirantes a policía municipal fueron seleccionados 100 basándose en las puntuaciones en tres test: X_1, X_2, X_3. Tras dos años ejerciendo su labor, se puntuó a los admitidos en dos criterios de eficacia laboral: Y_1 e Y_2. A partir de los datos obtenidos, y mediante las operaciones matriciales anteriormente indicadas, se obtuvieron las matrices de varianza-covarianza adjuntas para el grupo de aspirantes. En el supuesto de que hubiese que elegir un solo test para predecir el criterio Y_2, ¿cuál se elegiría?

$$C_{XY} \qquad C_{XX} \qquad C_{YY}$$

$$\begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array}\begin{array}{c} \begin{array}{cc} Y_1 & Y_2 \end{array} \\ \begin{bmatrix} 6 & 8 \\ 4 & 5 \\ 2 & 10 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array}\begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} 25 & & \\ 7 & 16 & \\ 6 & 4 & 9 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ Y_1 \\ Y_2 \end{array}\begin{array}{c} \begin{array}{cc} Y_1 & Y_2 \end{array} \\ \begin{bmatrix} 25 & \\ 11 & 16 \end{bmatrix} \end{array}$$

$$\rho_{X_1Y_2} = \frac{\text{cov}(X_1, Y_2)}{S_{X_1} S_{Y_2}} = \frac{8}{(5)(4)} = 0{,}40$$

$$\rho_{X_2Y_2} = \frac{\text{cov}(X_2, Y_2)}{S_{X_2} S_{Y_2}} = \frac{5}{(4)(4)} = 0{,}31$$

$$\rho_{X_3Y_2} = \frac{\text{cov}(X_3, Y_2)}{S_{X_3} S_{Y_2}} = \frac{10}{(3)(4)} = 0{,}83$$

Parece claro que el test 3 sería en solitario el mejor predictor, pues su correlación con el criterio 2 es bastante mayor que la de los otros dos test.

4. VALIDEZ Y PREDICCIÓN

Es frecuente que uno de los objetivos centrales de un test sea predecir determinado criterio o variable externa. La predicción será tanto más ajustada cuanta mayor correlación haya entre el test y el criterio a estimar, es decir, cuanto mayor sea el coeficiente de validez del test. A continuación se expondrá en líneas generales el modo de estimar un criterio mediante las técnicas de regresión, aconsejándose al lector acudir a las fuentes que se citan para tratamientos más detallados. Aunque conllevan cierta elaboración estadística, el lector debe tener presente a nivel conceptual que el dato fundamental en el que se basan es la correlación entre el test y el criterio. Puede decirse que las técnicas de regresión permiten expresar «de otro modo» la información contenida en el coeficiente de validez.

4.1. Regresión simple

Modelo

La regresión simple permite pronosticar un criterio a partir de un solo test. Desde el punto de vista de la psicología científica representaría un caso más bien atípico, en el sentido de que raramente una variable de interés (criterio) puede predecirse con precisión a partir de un solo test. Lo más realista suele ser que la predicción se ajuste mejor utilizando más variables predictoras; no obstante, tiene sentido tratarlo aquí aunque solo sea como introducción a la regresión múltiple.

Cuando se trata de pronosticar un criterio (Y) a partir de las puntuaciones conocidas de un test (X), la ecuación de la recta puede resultar un instrumento apropiado para tal fin, aunque otros muchos son pensables. La ecuación general de una recta viene dada por:

$$Y = bX + a$$

donde

b: Pendiente de la recta.
a: Ordenada en el origen.

Si se conocen los valores de b y a, la recta queda definida, y ciertamente se pueden pronosticar los valores de Y conocidos los de X. Por ejemplo, si en una recta $b = 2$ y $a = 3$, su fórmula vendría dada por $Y = 2X + 3$, así que a un valor de $X = 5$ le correspondería un valor de $Y = 13$; para un valor $X = 0$, Y valdría 3, etc. (véase figura 3.2).

Ahora bien, no es difícil demostrar (véase apéndice [2.8]) que los valores de b y a que hacen mínimos los errores al pronosticar Y a partir de X vienen dados por:

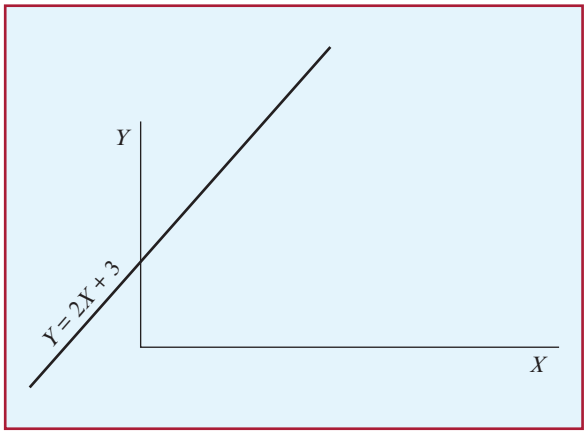


Figura 3.2.—Recta de regresión.

$$b = \rho_{xy}\frac{\sigma_y}{\sigma_x}$$

$$a = \bar{Y} - \rho_{xy}\frac{\sigma_y}{\sigma_x}\bar{X}$$

Sustituyendo en la recta los valores de *b* y *a:*

$$Y' = \rho_{xy}\frac{\sigma_y}{\sigma_x}X - \rho_{xy}\frac{\sigma_y}{\sigma_x}\bar{X} + \bar{Y}$$

Y sacando factor común $\rho_{xy}\frac{\sigma_y}{\sigma_x}$

$$Y' = \rho_{xy}\frac{\sigma_y}{\sigma_x}(X - \bar{X}) + \bar{Y} \qquad [3.18]$$

Ejemplo

Si el coeficiente de validez de un test es 0,80, la desviación típica del criterio 10, la del test 5, la media del criterio 60 y la del test 40, ¿qué puntuación se pronosticaría en el criterio a una persona que hubiese obtenido en el test una puntuación de 50?

$$Y' = (0{,}80)\frac{10}{5}(50 - 40) + 60 = 76$$

Por tanto, a las personas que hayan sacado 50 puntos en el test se les pronostican 76 en el criterio.

Error típico de estimación

Nótese bien que la recta de regresión lo único que garantiza es que «a la larga» (esperanza matemática) los errores cometidos al pronosticar serán mínimos, según el criterio de mínimos cuadráticos, pero es evidente que cada vez que se pronostique se cometerá algún error, mayor o menor, salvo que $\rho_{xy} = 1$. Estos errores cometidos al pronosticar se denominan «errores de estimación», y son la diferencia entre las puntuaciones pronosticadas en el criterio y las que realmente obtienen en él los sujetos $(Y' - Y)$, denominándose «error típico de estimación» a su desviación típica, cuya fórmula (véase apéndice [2.10]) viene dada por

$$\sigma_{y \cdot x} = \sigma_y\sqrt{1 - \rho_{xy}^2} \qquad [3.19]$$

donde

σ_y: Desviación típica del criterio.
ρ_{xy}: Coeficiente de validez del test.

Deducciones inmediatas del modelo

Es inmediato a partir del modelo (véase apéndice) que la varianza total del criterio (σ_y^2) puede desglosarse en dos componentes aditivos: la varianza de las puntuaciones pronosticadas en el criterio a partir del test $(\sigma_{y'}^2)$ y la varianza de los errores de estimación $(\sigma_{y \cdot x}^2)$:

$$\sigma_y^2 = \sigma_{y'}^2 + \sigma_{y \cdot x}^2 \qquad [3.20]$$

La varianza de las puntuaciones pronosticadas $(\sigma_{y'}^2)$ suele denominarse «varianza asociada», haciendo referencia a que, en efecto, lo que varían las puntuaciones pronosticadas Y' está asociada a, depende de la variabilidad de las puntuaciones X del test, ya que las Y' se obtienen a partir de las X mediante la recta de regresión. Por su parte, $\sigma_{y \cdot x}^2$ suele denominarse «varianza no asociada», pues, según el modelo, la varianza de los errores de estimación no depende de la variabilidad de las puntuaciones X.

Es también inmediato (véase apéndice) que:

$$\rho_{xy}^2 = \frac{\sigma_{y'}^2}{\sigma_y^2} \qquad [3.21]$$

$$\rho_{xy}^2 = 1 - \frac{\sigma_{y \cdot x}^2}{\sigma_y^2} \qquad [3.22]$$

La fórmula [3.21] muestra explícitamente que el coeficiente de validez al cuadrado expresa la proporción de varianza asociada entre el test y el criterio, o, en otras palabras, expresa qué proporción de la varianza del criterio se puede predecir a partir del test. Por ejemplo, si la validez de un test es 0,80, ello indicará que el 64% de la varianza del criterio es pronosticable a partir del test. Por ello, no es infrecuente denominar a ρ_{xy}^2 coeficiente de determinación y a $\sqrt{1-\rho_{xy}^2}$ coeficiente de alienación, aludiendo, respectivamente, al grado en que el criterio viene determinado por el test, o, por el contrario, está alienado, separado, enajenado del test. Nótese que el coeficiente de alienación viene dado por el cociente:

$$\frac{\sigma_{y \cdot x}}{\sigma_y} = \frac{\sigma_y \sqrt{1-\rho_{xy}^2}}{\sigma_y} = \sqrt{1-\rho_{xy}^2}$$

indicando la proporción que el error típico de estimación ($\sigma_{y \cdot x}$) representa respecto de σ_y.

Denomínase «coeficiente de valor predictivo» al complementario del coeficiente de alienación, $1-\sqrt{1-\rho_{xy}^2}$, otro modo de expresar la capacidad predictiva del test.

En suma, el coeficiente de validez y los índices citados derivados de él informan acerca del grado en que el criterio es pronosticable a partir del test.

— Coeficiente de determinación:

$$CD = \rho_{xy}^2$$

— Coeficiente de alienación:

$$CA = \sqrt{1-\rho_{xy}^2}$$

— Coeficiente de valor predictivo:

$$CVP = 1 - \sqrt{1-\rho_{xy}^2}$$

Intervalos confidenciales

A la hora de hacer pronósticos en el criterio a partir del test, y debido a los errores de estimación asociados con las predicciones, más que estimaciones puntuales conviene establecer un intervalo confidencial en torno a la puntuación pronosticada. Para ello se asume que los errores de estimación se distribuyen según la curva normal con desviación típica dada por el error típico de estimación ($\sigma_{y \cdot x}$).

Ejemplo

Se aplicó un test a una muestra de 100 personas, obteniéndose una media de 40 y una desviación típica de 5. La desviación típica de la muestra en el criterio fue 10, y la media, 60. El coeficiente de validez resultó ser de 0,90. Al nivel de confianza del 95%, ¿qué puntuación se estima que consiguiesen en el criterio los sujetos que obtuviesen 55 puntos en el test?

1. Nivel de confianza del 95%: $Z_c = \pm 1{,}96$.
2. $\sigma_{y \cdot x} = \sigma_y \sqrt{1-\rho_{xy}^2} = 10\sqrt{1-(0{,}90)^2} = 4{,}36$.
3. Error máximo: $(Z_c)(\sigma_{y \cdot x}) = (1{,}96)(4{,}36) = 8{,}54$.
4. $Y' = \rho_{xy} - (X - \bar{X}) + \bar{Y} = 0{,}90\,\frac{10}{5} - (55 - 40) + 60 = 87$.
5. $Y' \pm$Error máximo: $87 \pm 8{,}54$: $78{,}46 \leqslant Y \leqslant 95{,}54$.

Es decir, se estima que al nivel de confianza del 95% el valor de Y para las personas que obtuvieron 55 puntos en el test estará entre 78,46 y 95,54.

Siempre que las muestras sean suficientemente amplias, este modo de proceder es razonable; no obstante, véanse en la nota que sigue algunas matizaciones.

Aunque a estas alturas resulte obvio al lector, no conviene olvidar que la utilidad de los pronósti-

cos mediante la recta de regresión no tiene que ver con la muestra en la que se ha calculado, en la cual disponemos de las puntuaciones de los sujetos en el criterio, no teniendo, por tanto, ninguna necesidad de pronosticarlas; su utilidad proviene del uso que podamos hacer de ella en el futuro con sujetos equiparables a los que se emplearon en su elaboración.

NOTA: Para estimar el valor de $\sigma^2_{y\cdot x}$ a partir de los datos de una muestra, algunos autores, más que la fórmula dada en [3.19], especialmente si el número de sujetos no es elevado, aconsejan utilizar una corrección del estimador insesgado $\sigma'_{y\cdot x}$:

$$\sigma'_{y\cdot x} = S_{y\cdot x}\sqrt{\frac{N}{N-2}}$$

donde

$S_{y\cdot x}$: Valor de $\sigma_{y\cdot x}$ en la muestra.
N: Número de sujetos de la muestra.

Dicha corrección viene dada por:

$$\sigma''_{y\cdot x} = \sigma'_{y\cdot x}\sqrt{1+\frac{1}{N}+\frac{(X-\bar{X})^2}{(N-1)S_x^2}} \qquad [3.23]$$

donde

$\sigma'_{y\cdot x}$: Estimador insesgado citado.
N: Número de sujetos de la muestra.
X: Puntuación del test a pronosticar.
$\bar{X}$: Media del test en la muestra.
S^2_x: Varianza del test en la muestra.

La razón de recomendar el uso de esta corrección ($\sigma''_{y\cdot x}$) en vez de utilizar directamente $S_{y\cdot x}$ o el estimador insesgado $\sigma'_{y\cdot x}$ proviene de lo siguiente. Cuando se hacen pronósticos particulares de Y a partir de X, los errores de estimación tienden a ser mayores a medida que el valor de X se aleja de la media. Introduciendo la corrección citada, los intervalos confidenciales serán más amplios en los extremos, debido al factor $(X-\bar{X})^2$ que se contempla, y más estrechos para valores de X en torno a la media. Esto no ocurriría si se usase $S_{y\cdot x}$ o $\sigma'_{y\cdot x}$, que darían intervalos exactamente iguales independientemente del valor de X, lo cual se ajusta peor a lo que suele ocurrir empíricamente.

Además, el estadístico de contraste propuesto es t con $N-2$ grados de libertad en vez de Z.

Veamos lo dicho para los datos del ejemplo anterior:

1. Nivel de confianza del 95%: $t_{N-2} = \pm 1{,}984$.
2. $\sigma'_{y\cdot x} = 10\sqrt{1-0{,}90^2}\sqrt{\frac{100}{100-2}} = 4{,}40$

   $\sigma''_{y\cdot x} = 4{,}40\sqrt{1+\frac{1}{100}+\frac{(55-40)^2}{(100-1)25}} = 4{,}62.$
3. Error máximo: (1,984)(4,62) = 9,17.
4. $Y' = 87$.
5. $Y' \pm$ Error máximo: $87 \pm 9{,}17$: $77{,}83 \leqslant Y \leqslant 96{,}17$.

Dado que nuestro objetivo aquí no es hacer una exposición detallada de la regresión, sino más bien utilizar algunas de sus posibilidades para resolver problemas concretos de la validez de los test, no nos detendremos en aspectos como las distribuciones muestrales de los parámetros y en las distintas hipótesis acerca de ellos que se pueden someter a prueba. Dos de esas hipótesis de sumo interés son comprobar la significación estadística de la pendiente de la recta de regresión (H_0: $\beta = 0$) y el posible paralelismo entre dos rectas (H_0: $\beta_1 = \beta_2$). Véase, por ejemplo, Amón (1984).

A modo de ejercicio, demuestre el lector que en puntuaciones diferenciales la recta de regresión viene dada por:

$$y' = \rho_{xy}\frac{\sigma_y}{\sigma_x}x$$

mientras que en puntuaciones típicas

$$Z_{y'} = \rho_{xy}Z_x$$

4.2. Regresión múltiple

Si se desea pronosticar un criterio psicológico de cierta relevancia, lo más frecuente es que haya que utilizar más de una variable predictora. Por ejemplo,

un psicólogo escolar que esté interesado en predecir el rendimiento académico tal vez medirá distintos tipos de inteligencia (general, verbal, espacial, numérica o social), algunos rasgos de personalidad (extraversión, neuroticismo, motivación de logro, etc.), amén de otras variables como nivel socioeconómico familiar, procedencia rural-urbana o grado de motivación del profesorado, etc.; en suma, medirá aquellas variables que según los datos previos y la teoría con la que opera parezcan más relevantes. Ahora bien, no todas las variables en las que se pensó en principio serán igual de relevantes para la predicción, y tal vez algunas de ellas no contribuyan significativamente a la predicción. Pues bien, la regresión múltiple proporciona una solución plausible a esos problemas porque permite estimar los pesos o ponderaciones correspondientes a cada variable predictora, así como descartar aquellas cuya contribución a la predicción del criterio sea irrelevante. Veamos el modelo.

4.2.1. *Modelo*

Sean

Y: Criterio a predecir.
$X_1, X_2, X_3, ..., X_k$: Variables predictoras.
$B_1, B_2, B_3, ..., B_k$: Pesos o ponderaciones correspondientes a las variables predictoras.

Las puntuaciones pronosticadas en el criterio vendrán dadas por:

$$Y' = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \cdots + B_kX_k$$

Ahora bien, como ya se ha visto al tratar de la regresión simple, los pronósticos Y' no siempre coincidirán exactamente con el valor real de Y, cuya diferencia se denomina como allí «error de estimación»: $e = Y - Y'$. Por tanto, $Y = Y' + e$, pudiendo expresarse el modelo del siguiente modo:

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \cdots + B_kX_k + e$$

o, en forma matricial:

$$Y = X\mathbf{B} + \mathbf{e}$$

Estimación de los pesos de las variables predictoras

El problema consiste en cómo estimar a partir de los datos empíricos los pesos **B** para que los errores cometidos al pronosticar (e) sean mínimos. Se demuestra (véase apéndice) que los pesos de las variables predictoras que minimizan los errores de estimación según el criterio de mínimos cuadráticos vienen dados por:

$$\mathbf{b} = (X'X)^{-1}X'\mathbf{Y} \qquad [3.24]$$

donde

b: Vector de pesos estimados.
X: Matriz de sujetos por variables predictoras, cuya primera columna son unos.
X': Matriz traspuesta de X.
$(X'X)^{-1}$: Matriz inversa de $(X'X)$.
Y: Vector de puntuaciones de los sujetos en el criterio.

Nótese que para poder estimar los pesos **b** es necesario que la matriz $(X'X)$ tenga inversa. Una matriz cuyo determinante es cero no tiene inversa y se denomina «matriz singular», en cuyo caso no se podría hallar **b.** Aunque no es frecuente que ocurra esto con datos empíricos, si, por ejemplo, un investigador incluye en el análisis una variable que es función lineal de otra también incluida, se encontrará con la situación descrita.

Una vez estimados los pesos **b**, se pueden someter a prueba diferentes hipótesis estadísticas acerca de ellos, siendo de especial relevancia su significación estadística. A tal efecto, el vector de varianzas de **b** viene dado por $\sigma'^2_{y\cdot x}(X'X)^{-1}$, donde $\sigma'^2_{y\cdot x}$ es la varianza de los errores de estimación. Los programas de ordenador habituales, como el SPSS y otros, proporcionan este valor.

Dado que el objetivo que se persigue aquí con esta exposición sumaria de algunos aspectos de la regresión es permitir al lector entender los conceptos implicados en la teoría de los test, desbordando por completo los objetivos de este manual una exposición detallada, se recomienda al lector interesado acudir a los excelentes textos existentes al res-

pecto, por ejemplo Cohen y Cohen (1983), Draper y Smith (1981), Kerlinger y Pedhazur (1973), Overall y Klett (1972), Pedhazur (1982) o Timm (1975), entre otros.

Puntuaciones diferenciales

Si las puntuaciones directas de las personas en las variables predictoras y en el criterio se transformasen en puntuaciones diferenciales, el valor de $B_0 = 0$, y los pesos **b** vendrían dados por:

$$\mathbf{b} = C_{xx}^{-1} C_{xy} \qquad [3.25]$$

donde

C_{xx}^{-1}: Inversa de la matriz de varianzas-covarianzas de las variables predictoras.
C_{xy}: Vector de covarianzas entre las variables predictoras y el criterio.

Los pesos **b** en puntuaciones diferenciales son los mismos que en directas, excepto que $\mathbf{b_0} = 0$.

Puntuaciones típicas

En puntuaciones típicas los pesos de las variables predictoras suelen denominarse pesos *beta* y vienen dados por:

$$\boldsymbol{\beta} = R_{xx}^{-1} \mathbf{R}_{xy} \qquad [3.26]$$

donde

R_{xx}^{-1}: Inversa de la matriz de correlaciones (con unos en la diagonal) entre las variables predictoras.
$\mathbf{R}_{xy}$: Vector de correlaciones entre las variables predictoras y el criterio.

De cara a una mejor interpretación de la importancia relativa de las variables predictoras es aconsejable trabajar con los pesos *beta,* ya que todas las variables se expresan en la misma escala (típica), con media cero y desviación típica uno. No obstante, es inmediato el paso de **b** a **β**: $\beta_i = b_i S_{xi}/S_y$, es decir, el peso β de una variable i se obtiene multiplicando su peso b por su desviación típica y dividiendo por la desviación típica del criterio.

Ejemplo

Trataremos de ilustrar todo lo dicho mediante la realización de un ejemplo paso a paso. Supóngase que se desean conocer las ponderaciones de la inteligencia general (IG) y de la inteligencia verbal (IV) para pronosticar el rendimiento académico (RA). Con tal fin se midieron las tres variables en una muestra de cinco sujetos, obteniéndose los resultados que se adjuntan (tomados de Amón, 1984, p. 323).

Sujetos	Inteligencia general	Inteligencia verbal	Rendimiento académico
	X_1	X_2	Y
A	9	8	5
B	0	4	1
C	6	0	3
D	18	12	7
E	12	6	4

Identificaremos previamente la matriz X y el vector **Y**

$$X \qquad \mathbf{Y}$$
$$\begin{bmatrix} 1 & 9 & 8 \\ 1 & 0 & 4 \\ 1 & 6 & 0 \\ 1 & 18 & 12 \\ 1 & 12 & 6 \end{bmatrix} \quad \begin{bmatrix} 5 \\ 1 \\ 3 \\ 7 \\ 4 \end{bmatrix}$$

Nótese que la matriz X está formada por las puntuaciones de los sujetos en las variables predictoras precedidas por una columna de unos, y el vector **Y** son las puntuaciones de los sujetos en el criterio o variable a predecir.

Pesos b en puntuaciones directas

Para calcular los pesos **b** hay que ejecutar paso a paso las operaciones requeridas por la fórmula [3.24]:

$$\mathbf{b} = (X'X)^{-1} X'\mathbf{Y}$$

Paso 1. Hallar la traspuesta de X: X'.

Como es bien sabido, para encontrar la traspuesta de una matriz se convierten sus filas en columnas.

$$X'$$
$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 9 & 0 & 6 & 18 & 12 \\ 8 & 4 & 0 & 12 & 6 \end{bmatrix}$$

Paso 2. Efectuar el producto $X'X$.

$$\underset{X'}{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 9 & 0 & 6 & 18 & 12 \\ 8 & 4 & 0 & 12 & 6 \end{bmatrix}} \underset{X}{\begin{bmatrix} 1 & 9 & 8 \\ 1 & 0 & 4 \\ 1 & 6 & 0 \\ 1 & 18 & 12 \\ 1 & 12 & 6 \end{bmatrix}} = \underset{X'X}{\begin{bmatrix} 5 & 45 & 30 \\ 45 & 585 & 360 \\ 30 & 360 & 260 \end{bmatrix}}$$

Paso 3. Calcular la matriz inversa $(X'X)^{-1}$.

El cálculo de la inversa de una matriz requiere dividir la matriz traspuesta de los adjuntos entre el determinante:

$$A^{-1} = \frac{[\text{adj}(A)]'}{|A|}$$

Por tanto, es necesario el cálculo del determinante y de los adjuntos, lo cual realizaremos por medio de cinco subpasos.

1. Cálculo del determinante.

$$\begin{aligned} |A| &= (5)(585)(260) + (45)(360)(30) + \\ &+ (30)(45)(360) - (30)(585)(30) - \\ &- (5)(360)(360) - (45)(45)(260) = 31.500 \end{aligned}$$

2. Cálculo de los menores.

$a_{11} = (585)(260) - (360)(360) = 22.500$
$a_{12} = (45)(260) - (30)(360) = 900$
$a_{13} = (45)(360) - (30)(585) = -1.350$
$a_{21} = (45)(260) - (30)(360) = 900$
$a_{22} = (5)(260) - (30)(30) = 400$
$a_{23} = (5)(360) - (45)(30) = 450$
$a_{31} = (45)(360) - (30)(585) = -1.350$
$a_{32} = (5)(360) - (30)(45) = 450$
$a_{33} = (5)(585) - (45)(45) = 900$

3. Matriz de adjuntos.
   La matriz de adjuntos se obtiene a partir de los menores obtenidos en el paso anterior del siguiente modo: si los subíndices del menor suman impar, se le cambia el signo y si suman par se mantiene el que tiene.

$$[\text{adj}(A)]$$
$$\begin{bmatrix} 22.500 & -900 & -1.350 \\ -900 & 400 & -450 \\ -1.350 & -450 & 900 \end{bmatrix}$$

4. Traspuesta de la matriz de adjuntos.
   Será la misma del subpaso anterior 3, dado que la traspuesta de una matriz simétrica como esta es ella misma.
5. Se divide la traspuesta entre el determinante, obteniéndose $(X'X)^{-1}$.

$$(X'X)^{-1}$$
$$\begin{bmatrix} \frac{22.500}{31.500} & \frac{-900}{31.500} & \frac{-1.350}{31.500} \\ \frac{-900}{31.500} & \frac{400}{31.500} & \frac{-450}{31.500} \\ \frac{-1.350}{31.500} & \frac{-450}{31.500} & \frac{900}{31.500} \end{bmatrix}$$

Paso 4. Se multiplica X' por **Y.**

$$\underset{X'}{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 9 & 0 & 6 & 18 & 12 \\ 8 & 4 & 0 & 12 & 6 \end{bmatrix}} \underset{\mathbf{Y}}{\begin{bmatrix} 5 \\ 1 \\ 3 \\ 7 \\ 4 \end{bmatrix}} = \underset{X'\mathbf{Y}}{\begin{bmatrix} 20 \\ 237 \\ 152 \end{bmatrix}}$$

Paso 5. Se multiplica $(X'X)^{-1}$ por $(X'\mathbf{Y})$.

$$\underset{(X'X)^{-1}}{\begin{bmatrix} \frac{22.500}{31.500} & \frac{-900}{31.500} & \frac{-1.350}{31.500} \\ \frac{-900}{31.500} & \frac{400}{31.500} & \frac{-450}{31.500} \\ \frac{-1.350}{31.500} & \frac{-450}{31.500} & \frac{900}{31.500} \end{bmatrix}} \underset{X'\mathbf{Y}}{\begin{bmatrix} 20 \\ 237 \\ 152 \end{bmatrix}} = \underset{\mathbf{b}}{\begin{bmatrix} 0,99 \\ 0,26 \\ 0,10 \end{bmatrix}}$$

Por tanto, $b_0 = 0,99$, $b_1 = 0,26$ y $b_2 = 0,10$, pudiendo escribirse la ecuación de regresión en puntuaciones directas como sigue:

$$Y' = 0,99 + 0,26X_1 + 0,10X_2$$

Ecuación de regresión en puntuaciones diferenciales

Para obtener los pesos b en puntuaciones diferenciales seguimos los pasos indicados por la fórmula [3.25]:

$$\mathbf{B} = C_{xx}^{-1}C_{xy}$$

Las varianzas y covarianzas necesarias para configurar la matriz C_{xx} y el vector C_{xy} vienen dadas por:

$$\begin{aligned} \text{cov}\,(X_1, X_2) &= 18 \\ \text{cov}\,(X_1, Y) &= 11,4 \\ \text{cov}\,(X_2, Y) &= 6,4 \\ \text{var}\,(X_1) &= 36 \\ \text{var}\,(X_2) &= 16 \end{aligned}$$

$$\underset{C_{xx}}{\begin{bmatrix} 36 & 18 \\ 18 & 16 \end{bmatrix}} \quad \underset{C_{xy}}{\begin{bmatrix} 11,4 \\ 6,4 \end{bmatrix}}$$

Determinante de C_{xx}:

$$(36)(16) - (18)(18) = 252$$

$$\underset{C_{xx}^{-1}}{\begin{bmatrix} \frac{16}{252} & \frac{-18}{252} \\ \frac{-18}{252} & \frac{36}{252} \end{bmatrix}} \underset{C_{xy}}{\begin{bmatrix} 11,4 \\ 6,4 \end{bmatrix}} = \underset{\mathbf{b}}{\begin{bmatrix} 0,26 \\ 0,10 \end{bmatrix}}$$

Luego la ecuación de regresión en puntuaciones diferenciales viene dada por:

$$y' = 0,26x_1 + 0,10x_2$$

que es la misma que en directas sin el término independiente b_0.

Ecuación de regresión en puntuaciones típicas

La ecuación de regresión en puntuaciones típicas se obtiene según [3.26]:

$$\boldsymbol{\beta} = R_{xx}^{-1}\boldsymbol{R}_{xy}$$

Dado que la correlación entre X_1 y X_2 es 0,75; entre X_1 e Y, 0,95, y entre X_2 e Y, 0,80, se puede expresar en forma matricial:

$$\underset{R_{xy}}{\begin{bmatrix} 1,00 & 0,75 \\ 0,75 & 1,00 \end{bmatrix}} \quad \underset{\mathbf{R}_{xy}}{\begin{bmatrix} 0,95 \\ 0,80 \end{bmatrix}}$$

El determinante de R_{xx} será:

$$(1)(1) - (0,75)(0,75) = 0,4375$$

$$\underset{R_{xx}^{-1}}{\begin{bmatrix} \frac{1}{0,4375} & \frac{-0,75}{0,4375} \\ \frac{-0,75}{0,4375} & \frac{1}{0,4375} \end{bmatrix}} \underset{\mathbf{R}_{xy}}{\begin{bmatrix} 0,95 \\ 0,80 \end{bmatrix}} = \underset{\boldsymbol{\beta}}{\begin{bmatrix} 0,80 \\ 0,20 \end{bmatrix}}$$

Por tanto, la ecuación de regresión en puntuaciones típicas viene dada por la expresión:

$$Z_{y'} = 0,80Z_{x1} + 0,20Z_{x2}$$

Aunque más adelante se precisará esta afirmación, a la vista de la ecuación de regresión puede observarse que la variable X_1 tiene más peso (0,80), más importancia, a la hora de pronosticar Y que X_2 (0,20). En términos de nuestro ejemplo, la inteligencia general sería más importante que la inteligencia verbal para pronosticar el rendimiento académico.

Nótese que, una vez estimados los pesos de las variables predictoras, se pueden someter a prueba diferentes hipótesis acerca de ellos, según los objetivos del investigador, siendo de especial relevancia conocer si los pesos estimados resultan estadísticamente significativos para la predicción del criterio, que abordaremos al tratar de la correlación múltiple.

4.2.2. *Correlación múltiple*

Se entiende por correlación múltiple la correlación entre los pronósticos (Y') hechos a partir de la ecuación de regresión y el criterio (Y). $R_{y'y}$ indica, por tanto, en qué medida las variables predictoras tomadas conjuntamente permiten predecir el criterio. Elevada al cuadrado, la correlación múltiple expresa la proporción de varianza asociada (pronosticable) entre el criterio y las variables predictoras tomadas conjuntamente, es decir, $R^2_{y'} = S^2_{y'}/S^2_y$, donde $S^2_{y'}$ es la varianza asociada y S^2_y la varianza del criterio. Así, por ejemplo, si tuviésemos una correlación múltiple de 0,80, ello indicaría que el 64 % [$(0,80)^2 = 0,64$] de la varianza del criterio sería pronosticable a partir de las variables predictoras.

$R^2_{y'y}$ viene dado por la conocida fórmula (aquí en puntuaciones diferenciales) de la correlación: $R^2_{y'y} = (\Sigma y'y/NS_{y'}S_y)^2$, que, expresado en forma matricial, se convierte en:

$$R^2_{y'y} = \frac{\mathbf{b'C}_{xy}}{S^2_y} \qquad [3.27]$$

donde

- $\mathbf{b'}$: Vector traspuesto de los pesos de las variables predictoras.
- $\mathbf{C}_{xy}$: Vector de covarianzas entre el criterio y las variables predictoras.
- S^2_y: Varianza del criterio.

Calculemos la correlación múltiple para los datos del ejemplo del apartado anterior. Recuérdese que allí los pesos b eran, respectivamente, 0,26 y 0,10, las covarianzas eran 11,4 y 6,4 y la varianza de Y era 4, es decir:

$$\mathbf{b'} = \begin{bmatrix} 0,26 & 0,10 \end{bmatrix} \; ; \; \mathbf{C}_{xy} = \begin{bmatrix} 11,4 \\ 6,4 \end{bmatrix} \; ; \; S^2_y = 4$$

Por tanto, aplicando [3.27]:

$$R^2_{y'y} = \frac{\begin{bmatrix} 0,26 & 0,10 \end{bmatrix}\begin{bmatrix} 11,4 \\ 6,4 \end{bmatrix}}{4} = 0,9$$

La correlación múltiple al cuadrado es 0,9, que, en términos de nuestro problema, significaría que el 90 % de la varianza del rendimiento académico es pronosticable a partir de la inteligencia general y de la inteligencia verbal tomadas conjuntamente.

La correlación múltiple puede expresarse también en términos de los pesos $\boldsymbol{\beta}$:

$$R^2_{y'y} = \boldsymbol{\beta}\mathbf{R}_{xy} \qquad [3.28]$$

donde

- $\boldsymbol{\beta}$: Vector de pesos β traspuesto.
- $\mathbf{R}_{xy}$: Vector de correlaciones entre el criterio y las variables predictoras.

Que aplicado a los datos de nuestro ejemplo arroja obviamente los mismos resultados:

$$R^2_{y'y} = \begin{bmatrix} 0,80 & 0,20 \end{bmatrix}\begin{bmatrix} 0,95 \\ 0,80 \end{bmatrix} = 0,9$$

Finalmente, en puntuaciones directas $R^2_{y'y}$ viene dado por:

$$R^2_{y'y} = \frac{b'X'Y - \left(\sum_{i=1}^{N} Y_i\right)^2 / N}{Y'Y - \left(\sum_{i=1}^{N} Y_i\right)^2 / N}$$

que para nuestros datos:

$$R^2_{y'y} = \frac{98 - 80}{100 - 80} = 0{,}9$$

a) *Estimador insesgado de* $R^2_{y'y}$

$$\rho^2_{y'y} = 1 - \frac{(N-1)(1-R^2_{y'y})}{N-K-1} \qquad [3.29]$$

donde

N: Número de sujetos de la muestra.
K: Número de variables predictoras.
$R_{y'y}$: Correlación múltiple en la muestra.

A esta corrección de $R_{y'y}$ se la denomina a veces «ajuste», por ejemplo, en el paquete de programas estadísticos SPSS.

b) *Comprobación de hipótesis acerca de la correlación múltiple*

Existen dos tipos de hipótesis de especial interés acerca de la correlación múltiple:

1. Significación estadística de la correlación.
2. Significación estadística de las diferencias.

*b.*1. *Significación estadística de la correlación*

Tal vez la primera pregunta tras calcular la correlación múltiple en una muestra sea si esta es estadísticamente significativa, es decir, si el valor hallado es compatible con la hipótesis de que el verdadero valor en la población sea cero:

$$H_0\colon \rho^2_{y'y} = 0$$
$$H_1\colon \rho^2_{y'y} \neq 0$$

En cuyo caso el estadístico de contraste viene dado por:

$$F = \left(\frac{N-K-1}{K}\right)\left(\frac{R^2_{y'y}}{1-R^2_{y'y}}\right) \qquad [3.30]$$

que se distribuye según F con K y $(N - K - 1)$ grados de libertad.

Ejemplo

En una muestra de 100 personas se encontró una correlación múltiple de 0,60 para una ecuación de regresión con cinco variables predictoras. Al nivel de confianza del 95%, ¿puede afirmarse que la correlación hallada es estadísticamente significativa?

$$F = \left(\frac{100-5-1}{5}\right)\left(\frac{0{,}60^2}{1-0{,}60^2}\right) = 10{,}575$$

Dado que el valor crítico de F en las tablas correspondientes con 5 y 94 grados de libertad es 2,67, menor que 10,575, rechazamos la hipótesis nula, afirmando que la correlación múltiple es estadísticamente significativa al nivel de confianza del 95%. Si además de estadísticamente significativa es psicológicamente relevante, es cuestión que el investigador tendrá que indagar y decidir allegando más datos y contemplándola a la luz de su marco teórico de referencia.

*b.*2. *Significación estadística de las diferencias*

La segunda hipótesis de interés se refiere a la significación estadística de la diferencia entre dos correlaciones múltiples en la situación que se describe a continuación. Todo investigador intenta encontrar modelos lo más parsimoniosos posible; así, por ejemplo, si ha construido un modelo de regresión con cinco variables predictoras, podría plantearse si ello supondrá alguna mejora sensible respecto a tener solo tres de ellas. En otras palabras, se plantea si un modelo más parsimonioso, con solo tres predictores, es igualmente eficaz para pronosticar el criterio. Una estrategia posible para abordar esa situación será calcular la correlación múltiple de las cinco variables con el criterio, calcularla asimismo con tres y ver si la diferencia resulta estadísticamente significativa. Si no resultase, sería legítimo estadísticamente utilizar el modelo de tres predictores en vez del de cinco.

El estadístico de contraste que permite someter a pruebas la citada hipótesis:

$$H_0\colon \rho^2_{Y'_kY} - \rho^2_{Y'_pY} = 0 \quad , \quad p < k$$

viene dado por:

$$F = \left(\frac{N-k-1}{k-p}\right)\left(\frac{R^2_{Y'_kY} - R^2_{Y'_pY}}{1-R^2_{Y'_kY}}\right) \qquad [3.31]$$

que se distribuye según F con $(k-p)$ y $(N-k-1)$ grados de libertad y donde:

N: Número de sujetos.
k y p: Número de predictores con $p < k$.
$R_{y'_ky}R_{y'_py}$: Correlaciones múltiples con k y p predictores, respectivamente.

Ejemplo

En una muestra de 100 personas, la correlación múltiple (al cuadrado) de seis variables predictoras con cierto criterio fue de 0,80 ($R^2_{y'_6y} = 0{,}80$). La correlación con ese mismo criterio pero utilizando solo cuatro de las variables predictoras anteriores resultó ser 0,78 ($R^2_{y'_4y} = 0{,}78$). Al nivel de confianza del 95%, ¿pueden eliminarse del modelo las variables predictoras 5 y 6 sin una pérdida sustancial de su capacidad predictiva?

$$H_0\colon \rho^2_{y'_6y} - \rho^2_{y'_4y} = 0$$

$$F = \left(\frac{100-6-1}{6-4}\right)\left(\frac{0{,}80-0{,}78}{1-0{,}80}\right) = 4{,}65$$

En las tablas el valor crítico de F con 2 y 93 grados de libertad es 3,07, menor que 4,65; luego, como la diferencia resulta estadísticamente significativa, se rechaza la hipótesis nula, lo que indicaría que las variables predictoras 5 y 6 contribuyen significativamente al pronóstico del criterio. Nótese, no obstante, que su inclusión solo aumenta en dos centésimas el valor de la correlación múltiple al cuadrado (0,80 − 0,78 = 0,02), lo que da pie para plantearse la recurrente polémica de la significación estadística versus la significación psicológica.

Error típico de estimación

Como ya se ha indicado al tratar la regresión simple, en el modelo de regresión los pronósticos no coincidirán exactamente con las puntuaciones reales en el criterio. Lo único que garantiza el modelo es que los errores cometidos serán mínimos, según el criterio de mínimos cuadráticos. A los errores cometidos al pronosticar, es decir, a la diferencia $(Y' - Y)$, se les denomina «errores de estimación», y a su desviación típica ($S_{y\cdot x}$), «error típico de estimación»:

$$S_{y\cdot x} = \sqrt{\frac{\sum (Y'-Y)^2}{N}} \qquad [3.32]$$

La fórmula [3.32] refleja directamente el concepto del error típico de estimación, aunque su cálculo suele hacerse más bien en términos matriciales:

$$S_{y\cdot x} = \sqrt{\frac{\mathbf{Y}^t\mathbf{Y} - \mathbf{b}'X'\mathbf{Y}}{N}} \qquad [3.33]$$

donde todos los términos son los citados al exponer el modelo de regresión, con $\mathbf{Y}^t$ como vector traspuesto de $\mathbf{Y}$, usándose aquí la t para evitar su confusión con los pronósticos (Y').

Si se dispone de la correlación múltiple, hecho habitual, el cálculo más sencillo de $S_{y\cdot x}$ se realiza mediante:

$$S_{y\cdot x} = S_y\sqrt{1-R^2_{y'_py}} \qquad [3.34]$$

Un estimador insesgado del valor de $S_{y\cdot x}$ en la población viene dado por:

$$S^*_{y\cdot x} = S_{y\cdot x}\sqrt{\frac{N}{N-k-1}} \qquad [3.35]$$

siendo N el número de sujetos de la muestra y k el número de variables predictoras.

El error típico de estimación es otra forma de expresar el grado de ajuste entre el criterio y los pronósticos hechos a partir de las variables predictoras mediante la ecuación de regresión. Si todos los pronósticos coincidiesen exactamente con los valores reales del criterio, entonces $Y' - Y = 0$. No habría errores de estimación; por ende, su desvia-

ción típica (el error típico de estimación) también sería 0, o, lo que es lo mismo, la correlación múltiple sería 1. Valiéndose de $S^*_{y \cdot x}$ pueden establecerse intervalos confidenciales en torno a Y', utilizando t con $N - k - 1$ grados de libertad.

4.2.3. *Correlación parcial y semiparcial*

Existen numerosas situaciones en las que los investigadores están interesados en el estudio de la relación entre dos variables controlando el influjo de una tercera, o de varias. Cuando ello es posible, el mejor método de control es el experimental. Por ejemplo, si se está investigando la relación entre la inteligencia y el rendimiento académico, y se quiere controlar, mantener fija, la motivación, que lógicamente puede estar afectando a dicha relación, la mejor forma de hacerlo sería trabajar con sujetos todos ellos con la misma motivación (control experimental). Ahora bien, en psicología eso no siempre es posible, fácil casi nunca, por lo que a veces el único control posible es el control estadístico mediante la correlación parcial, la cual permite, en ciertos supuestos, calcular la correlación entre dos variables eliminando el influjo de una tercera, o, en el caso multivariado, el influjo de *n*.

La lógica del método es la siguiente. Supongamos que se desea correlacionar el criterio Y con la variable X, eliminando el influjo de Z. Pues bien, mediante las ecuaciones de regresión correspondientes se pronostican X e Y a partir de Z. La correlación parcial es la correlación entre las diferencias $(Y' - Y)$ y $(X' - X)$, es decir, es la correlación entre X e Y una vez que se les ha restado a ambas aquello que tienen de pronosticable a partir de Z. Nótese que al restar de X y de Y lo pronosticable a partir de Z, lo que queda, las diferencias, es lo que no depende de Z. Si se desea controlar el efecto no solo de una variable, como en este caso, sino de varias tomadas conjuntamente, la lógica es exactamente la misma; lo único que varía es que para pronosticar X e Y en vez de una recta de regresión habrá que utilizar una ecuación de regresión con aquellas n variables cuyo influjo se desea controlar.

La fórmula general para el cálculo de la correlación parcial vendría dada por la fórmula general de la correlación de Pearson en puntuaciones diferenciales aplicada a las restas citadas, es decir:

$$r_{xy \cdot z} = \frac{\sum (y' - y)(x' - x)}{N S_{(y' - y)} S_{(x' - x)}} \qquad [3.36]$$

donde x' e y' son los pronósticos de las variables x e y a partir del vector de variables **Z** cuyo efecto se quiere eliminar, N es el número de sujetos, $S_{(y' - y)}$ y $S_{(x' - x)}$ son las desviaciones típicas de los errores de estimación correspondientes, o errores típicos de estimación ($S_{y \cdot z}$, $S_{x \cdot z}$).

En el caso de solo tres variables, todos los pasos anteriores pueden resumirse en la siguiente expresión:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{zy} r_{zx}}{\sqrt{1 - r^2_{zx}} \sqrt{1 - r^2_{zy}}} \qquad [3.37]$$

donde X e Y son las variables a correlacionar y Z la variable cuyo efecto se desea eliminar.

Ejemplo

En una muestra de escolares se obtuvo una correlación entre inteligencia (X) y rendimiento académico (Y) de 0,60. A su vez, la correlación entre la inteligencia y la motivación de logro (Z) fue de 0,40, y entre la motivación y el rendimiento, de 0,80. ¿Cuál sería la correlación entre el rendimiento y la inteligencia si se eliminase el influjo de la motivación de logro?

$$r_{xy \cdot z} = \frac{0{,}60 - (0{,}80)(0{,}40)}{\sqrt{1 - 0{,}40^2} \sqrt{1 - 0{,}80^2}} = 0{,}50$$

Al eliminar el efecto de la motivación de logro, la correlación entre el rendimiento y la inteligencia baja de 0,60 a 0,50.

Correlación semiparcial

Como se acaba de señalar, en la correlación parcial entre dos variables se elimina el influjo de una tercera, o de varias, sobre las dos que se correlacionan. Pues bien, en el caso de la correlación semiparcial, como su nombre indica, solo se elimina el influjo sobre una de ellas. La lógica es exacta-

mente igual que en la correlación parcial, pero aquí solo se ejerce el control sobre una de las variables correlacionadas. Por ejemplo, y utilizando la terminología anterior, la correlación semiparcial entre X e Y controlando el efecto de Z sobre X vendría dada por:

$$r_{(x-x')y} = \frac{\sum (x - x')y}{NS_{(x-x')}S_y} \qquad [3.38]$$

donde x' son los pronósticos de x a partir del vector de variables **Z** cuyo efecto se desea parcializar.

Para el caso de solo tres variables [3.38] puede expresarse:

$$r_{(x-x')y} = \frac{r_{xy} - r_{zx}r_{zy}}{\sqrt{1 - r_{zx}^2}} \qquad [3.39]$$

Para el ejemplo anterior, si solo se controlase el efecto de la motivación de logro (Z) sobre la inteligencia (X), la correlación semiparcial entre inteligencia y rendimiento sería:

$$r_{(x-x')y} = \frac{0{,}60 - (0{,}40)(0{,}80)}{\sqrt{1 - 0{,}40^2}} = 0{,}30$$

El concepto de correlación semiparcial es de suma importancia para entender cabalmente la regresión y correlación múltiples. Así, lo que aumenta la correlación múltiple al añadir una variable predictora a la ecuación de regresión es precisamente la correlación semiparcial de esa variable predictora añadida con el criterio, parcializando el influjo sobre ella de las otras variables predictoras que ya estaban incluidas en el modelo. En la técnica del análisis de covarianza también se utiliza la lógica de la correlación semiparcial.

Selección de la mejor ecuación de regresión

Como acabamos de ver en los apartados precedentes, el modelo de regresión múltiple permite estimar los pesos de las variables predictoras que el investigador ha incluido en el análisis, pero en muchas ocasiones resulta que no todas las variables incluidas son relevantes para la predicción del criterio. Existen diversos métodos estadísticos para descartar las variables no significativas para la predicción, y una buena revisión puede consultarse en Hocking (1976), Younger (1979) o Draper y Smith (1981). Uno de los métodos más utilizados es el *stepwise* (paso a paso), implementado en la mayoría de los programas de ordenador. La lógica general del método, que no su exposición detallada (acúdase para ello a la literatura especializada citada), consiste en lo siguiente.

Supóngase, por ejemplo, que un investigador ha utilizado K variables predictoras y desea saber si puede disponer de un modelo más parsimonioso (que incluya menos variables predictoras), que explique un porcentaje de varianza del criterio similar al explicado por los K predictores. Pues bien, el método *stepwise* empieza cogiendo una de esas variables predictoras y va añadiendo las otras, una a una, paso a paso. ¿Con qué criterio lo hace? ¿Cuándo se para? La primera que elige, lógicamente, es la que mayor correlación tenga con el criterio. La segunda será aquella cuya correlación semiparcial con el criterio (parcializando el influjo de la ya admitida en primer lugar) sea más elevada. En tercer lugar, incluirá aquella cuya correlación semiparcial con el criterio (eliminando el influjo de las dos ya admitidas) sea mayor, y así sucesivamente. En cada uno de estos pasos se reanalizan las variables incluidas y eventualmente alguna de ellas puede descartarse de nuevo. El proceso de entrada de nuevas variables se detiene cuando al añadir otra la correlación múltiple con el criterio no se incrementa significativamente, es decir, la nueva variable no añade información significativa sobre el criterio respecto de la aportada por las ya incluidas. La significación estadística de esa diferencia puede evaluarse mediante el estadístico de contraste propuesto en [3.31]. Los paquetes estadísticos como el SPSS y otros muchos ofrecen salidas con el comportamiento estadístico de las variables en cada paso.

El método *stepwise* es del tipo *forward* (hacia adelante), pues empieza seleccionando una variable y continúa añadiendo otras según la lógica expuesta. Hay otros métodos de tipo *backward* (hacia atrás), que empiezan incluyendo todas las variables predictoras en la ecuación y luego van descartando las menos relevantes.

En general, si bien los diferentes métodos se prestan a interesantes y justificadas polémicas esta-

dísticas, a nivel empírico tienden a converger razonablemente.

Validez incremental

Suele denominarse «validez incremental de una variable» al aumento experimentado por la correlación múltiple al incluir dicha variable en el modelo de regresión.

Validez cruzada

Los pesos de las variables predictoras se estiman en una muestra de personas determinada. Cabe, pues, preguntarse si esas estimaciones son invariantes, es decir, si serían las mismas calculadas en una muestra diferente de sujetos. Dícese haber validez cruzada cuando se da dicha invarianza. Una forma de saberlo empíricamente es elaborar la ecuación de regresión en diferentes muestras y luego comparar los pesos y correlaciones múltiples obtenidos. Las diferencias entre las muestras no han de ser estadísticamente significativas. Nótese que no es infrecuente cuando las variables predictoras están muy intercorrelacionadas (multicolinealidad) obtener correlaciones múltiples similares con pesos bien distintos para las variables predictoras.

4.2.4. *Variables moduladoras y supresoras*

Se denomina «variables moduladoras» (Saunders, 1956) a aquellas que modulan, afectan, modifican la validez. Por ejemplo, una ecuación de regresión para predecir el rendimiento académico a partir de ciertos test puede tener diferente eficacia (validez diferencial) según los sujetos sean de clase social alta o baja, es decir, su validez viene modulada por la variable clase social. La literatura psicológica está plagada de ejemplos al respecto; véase, por ejemplo, Anastasi (1988). En estas circunstancias, o bien se utilizan ecuaciones de regresión distintas de acuerdo con la variable moduladora, en nuestro caso una ecuación para la clase social alta y otra para la baja, o en algunos casos sería posible elaborar una ecuación de regresión no lineal, hoy factible gracias a los ordenadores. Un buen análisis sobre las variables moduladoras puede consultarse en Zedeck (1971), y una discusión actualizada de algunos problemas estadísticos implicados, en Bobko (1986), Cronbach (1987), Dunlap y Kemery (1987), Lubinski y Humphreys (1990), McClelland y Judd (1993) o Morris et al. (1986).

Variables supresoras

Se denominan así aquellas variables que, paradójicamente, aun sin correlacionar con el criterio, si se las incluye en la ecuación de regresión mejoran la precisión de los pronósticos, la validez predictiva. La explicación radica en las correlaciones de estas variables con las otras predictoras incluidas en la ecuación de regresión, a las que añaden varianza irrelevante que reduce su correlación con el criterio. De ahí que si se incluyen estas variables supresoras en la ecuación de regresión se elimine esa varianza indeseada, potenciándose la relación de las otras con el criterio. Horst (1966) encontró, por ejemplo, que la inteligencia mecánica, la inteligencia numérica y la inteligencia espacial eran buenos predictores del éxito pilotando un avión, pero no así la inteligencia verbal. Sin embargo, al incluir esta en la ecuación de regresión, mejoraban las predicciones. La explicación más plausible es que la inteligencia verbal correlacionaba bastante con los otros tres tipos de inteligencia, y al incluirla en la ecuación permitía descontar aquella parte de las puntuaciones de los sujetos en inteligencia mecánica, numérica y espacial debida al mero hecho de tener una buena inteligencia verbal, con lo que se «purificaban de verbo» los predictores, mejorando los pronósticos. Aunque interesantes a nivel conceptual, en la literatura empírica no es frecuente la presencia de variables supresoras. Análisis detallados de la problemática implicada en este tipo de variables pueden consultarse en Conger (1974) o Tzelgov y Stern (1978).

4.3. Validez y decisión

En numerosas situaciones aplicadas el criterio que se pretende predecir solo toma dos valores, por ejemplo, aprobar-suspender, enfermo-no enfermo, admitido-rechazado, etc. En estos casos tiene poco sentido establecer la validez de la prueba mediante un coeficiente de correlación, aunque se puede hacer, pero se trata más bien de comprobar si las cla-

sificaciones hechas por el test son las adecuadas, es decir, si las decisiones tomadas a partir de las puntuaciones de las personas en el test coinciden con las del criterio. El caso más habitual es que los jueces clasifiquen a las personas en dos categorías y las puntuaciones en el test predictor se dicotomicen por determinado punto de corte, obteniéndose una tabla de contingencia de 2 × 2. En esta situación, si el test fuese perfectamente válido, las clasificaciones hechas a partir de las puntuaciones de las personas en él serían idénticas a las realizadas por los expertos (criterio). Por tanto, en este contexto el estudio de la validez de las pruebas se refiere al análisis de la convergencia entre las decisiones tomadas a partir de la prueba y las del criterio, habitualmente expertos, aunque otras opciones son posibles. Esta concordancia o discordancia clasificatoria puede evaluarse mediante varios índices y estrategias que se expondrán a continuación.

4.3.1. *Índices de validez*

Veamos la lógica de los distintos índices propuestos mediante un ejemplo numérico. Sea una muestra de 12 personas que han sido diagnosticadas (criterio) por un equipo de psicólogos en dos grupos: las que necesitan terapia antidepresiva (*TE*) y las que no la necesitaban (*NT*). A esas mismas 12 personas se les aplicó una escala de depresión de 10 puntos, obteniéndose los resultados de la tabla adjunta:

Personas	Escala	Diagnóstico
A	6	*NT*
B	6	*TE*
C	7	*TE*
D	8	*NT*
E	5	*NT*
F	8	*TE*
G	4	*NT*
H	9	*TE*
I	3	*NT*
J	7	*TE*
K	7	*NT*
L	10	*TE*

Se considera que toda persona que obtenga 7 puntos o más en la escala de depresión necesita terapia. ¿Cuál es la validez de la escala para predecir las personas que necesitan terapia?

Para responder a esta pregunta podría procederse tal como se indicó hasta ahora para estimar el coeficiente de validez, es decir, calculando la correlación entre la escala y los diagnósticos (criterio). Nada lo impide, y sería correcto, se obtendría un indicador de la validez de la escala, pero resultaría demasiado general; por ejemplo, no informaría de los distintos tipos de errores cometidos al clasificar mediante la escala, lo cual es muy importante, pues, como luego se verá, no todos los tipos de errores tienen la misma importancia en todas las situaciones. En este contexto, más que mediante un coeficiente de correlación, la validez se va a analizar en función de la coincidencia entre las decisiones hechas a partir del test y las obtenidas en el criterio. Es importante entender que ambos conceptos no se oponen, convergen, pero el estudio de la concordancia de las decisiones permite un análisis más detallado que el mero cálculo de la correlación.

Para responder a la pregunta del ejemplo, lo primero que hay que hacer es elaborar una tabla de contingencia de 2 × 2 en la que se reflejen las decisiones hechas a partir de la escala una vez fijado el punto de corte. Compruebe el lector que, a partir de los datos del ejemplo con el punto de corte establecido en 7 puntos, se generaría la siguiente tabla:

		Escala	
		No terapia	**Terapia**
Diagnóstico	Terapia	1	*5*
	No terapia	4	2

Como se puede observar, los datos de la tabla se reparten del siguiente modo:

- *Falsos positivos:* 2. Son las dos personas (*D* y *K*) que la escala detecta como necesitadas de terapia, mientras que los expertos (criterio) consideran que no la necesitan.
- *Falsos negativos:* 1. Según la escala, la persona *B* no necesita terapia; sin embargo, los expertos consideran que sí.

— *Aciertos:* 9. Son las personas correctamente clasificadas, cinco que según la escala requieren terapia, y así es según los expertos, y cuatro que la escala predice que no la necesitan, coincidiendo en ello con los expertos.

A partir de esos datos pueden obtenerse diversos indicadores de la validez de la escala para pronosticar el criterio:

Proporción de clasificaciones correctas

Es la proporción de clasificaciones correctas hechas a partir del test. Para nuestros datos vendría dado por:

$$P_c = \frac{5+4}{12} = 0{,}75$$

Sensibilidad

Proporción de personas correctamente detectadas por la escala respecto del total de casos existentes según los expertos.

En nuestro ejemplo, según los expertos hay seis personas que necesitarían terapia, de las cuales cinco son correctamente detectadas por la escala. Por tanto, la sensibilidad de la escala viene dada por:

$$S = \frac{5}{6} = 0{,}83$$

La sensibilidad será máxima cuando no haya falsos negativos.

Especificidad

Proporción de personas correctamente consideradas por la escala como no necesitadas de terapia respecto del total de personas que según los expertos no necesitan terapia.

De las seis personas que según los expertos no necesitan terapia, nuestra escala identifica correctamente a cuatro; por tanto, la especificidad viene dada por:

$$E = \frac{4}{6} = 0{,}67$$

La especificidad será máxima cuando no existan falsos positivos.

Nótese que, según las circunstancias de cada situación, se puede desear maximizar o bien la sensibilidad, o bien la especificidad, idealmente ambas, claro. Con frecuencia ambos índices se utilizan en porcentajes, multiplicando el valor obtenido por 100.

Coeficiente *kappa*

Este coeficiente, expuesto al tratar la fiabilidad de los test referidos al criterio, ofrece un indicador general de la validez de las clasificaciones hechas por el test. Su fórmula (véase el subepígrafe 9.2 del capítulo 2) viene dada por:

$$K = \frac{F_c - F_a}{N - F_a}$$

Aplicando la fórmula a los datos de la tabla de contingencia del ejemplo:

$$F_c = 5 + 4 = 9$$

$$F_a = \frac{6 \times 7}{12} + \frac{6 \times 5}{12} = 6$$

$$K = \frac{9 - 6}{12 - 6} = 0{,}5$$

Puesto que el valor máximo del coeficiente *kappa* es 1, la validez de la escala como predictor de las clasificaciones de los expertos es moderada. Parece contrastar este dato con la proporción de aciertos totales, que era 0,75, pero no se olvide que en ese valor de 0,75 están incluidas las clasificaciones correctas debidas al mero azar; lo que hace el coeficiente *kappa* es corregir esa anomalía, restando de los aciertos aquellos debidos al azar.

La utilización de un indicador u otro depende de los intereses del investigador y profesional en cada momento y del tipo de errores que se pretenda minimizar (véase el cuadro «Índices de validez de decisión» de la página siguiente).

Índices de validez de decisión

		Test	
		No caso	Caso
Criterio	Caso	*a*	*b*
	No caso	*c*	*d*

a: Falsos negativos.
b: Verdaderos positivos.
c: Verdaderos negativos.
d: Falsos positivos.

$$\text{Aciertos totales: } \frac{b+c}{a+b+c+d}$$

$$\text{Sensibilidad: } \frac{b}{a+b}$$

$$\text{Especificidad: } \frac{c}{c+d}$$

$$\textit{Kappa}\text{: } K = \frac{F_c - F_a}{N - F_a}$$

$$\text{donde: } F_c = b + c$$

$$F_a = \frac{(a+c)(c+d)+(b+d)(a+b)}{a+b+c+d}$$

$$N = a + b + c + d$$

4.3.2. *Incidencia del punto de corte en los tipos de errores*

Veamos con algo más de detenimiento los distintos tipos de errores en función de los puntos de corte establecidos. En el ejemplo de la escala de depresión se estableció que se consideraría necesitadas de tratamiento a todas aquellas personas que obtuviesen 7 puntos o más. Pero, ¿por qué 7 puntos y no otro valor? ¿Qué valor del punto de corte hace más precisas las clasificaciones de la escala? Este es un asunto importante, la validez de la escala va a depender del punto de corte, pero no solo de eso, sino también de la importancia que demos a los dos tipos de errores cometidos: falsos positivos y falsos negativos. A continuación se ofrecen los errores cometidos al considerar todos los puntos de corte posibles, desde 1 hasta 10. Compruebe el lector los resultados, obtenidos a partir de los datos originarios del ejemplo:

Punto de corte	Falsos positivos	Falsos negativos	Total errores
1	6	0	6
2	6	0	6
3	6	0	6
4	5	0	5
5	4	0	4
6	3	0	3
7	2	1	3
8	1	3	4
9	0	4	4
10	0	5	5

A la vista de estos datos, la pregunta es: ¿dónde resulta más eficaz ubicar el punto de corte?; en otras palabras, ¿qué punto de corte maximiza las clasificaciones correctas? Bien, la respuesta no es directa sin más. Por un lado, parece claro que el punto de corte con el que menos errores totales se cometen es o bien 6 (tres errores) o bien 7 (tres errores). Debería dar igual elegir el 6 o el 7, pero si nos fijamos un poco más, se observa que los tres errores del 6 provienen los tres de falsos positivos, mientras que con respecto a los del 7, dos provienen de falsos positivos y uno de falsos negativos. De modo que la elección de 6 o 7 dependerá de cómo valoremos los dos tipos de errores. Si, por ejemplo, la terapia antidepresiva es barata y carece de efectos secundarios, podría ser más grave dejar a alguien que la necesita sin ella (falsos negativos) que administrarla a alguien que no la necesitase (falsos positivos). Si este fuera el caso, elegiríamos como punto de corte 6 frente a 7. Por ejemplo, imaginemos que el equipo de profesionales considera el triple de graves los falsos negativos que los falsos positivos. Bajo esta ponderación, recalculamos todos los errores totales (T'), multiplicando los falsos negativos por 3:

Punto de corte	Errores totales ponderados
1	6 + 3(0) = 6
2	6 + 3(0) = 6
3	6 + 3(0) = 6
4	5 + 3(0) = 5
5	4 + 3(0) = 4
6	**3 + 3(0) = 3**0
7	2 + 3(1) = 5
8	1 + 3(3) = 10
9	0 + 3(4) = 12
10	0 + 3(5) = 15

Es evidente que bajo ese criterio establecer el punto de corte en 7 es más gravoso (cinco errores) que hacerlo en 6 (tres errores). Muchas situaciones son pensables, dependerá en cada caso particular el que unos errores sean más importantes que otros. Saberlo es importante, vital, para el establecimiento del punto de corte. Obsérvese, por ejemplo, cómo para los datos anteriores sin ponderar el número de errores totales es el mismo cuando se establece el punto de corte en 4 que cuando se establece en 10; sin embargo, la naturaleza de los errores es justamente la contraria; si nos fijásemos únicamente en el número total de errores, podríamos decir que da igual ubicar el punto de corte en 4 que en 10, lo cual puede ser incorrecto en muchas situaciones. En términos de la teoría de la decisión, ello quiere decir que siempre hay que tener en cuenta la matriz de pagos, es decir, las penalizaciones correspondientes a los distintos tipos de errores, en relación con los beneficios de los aciertos. En unos casos interesará minimizar un tipo de errores y en otros casos tal vez otros.

4.3.3. *Curvas ROC*

Hemos visto en el apartado anterior la incidencia de los puntos de corte sobre los errores cometidos por un instrumento de medida al realizar clasificaciones. Una forma sistemática de analizar el funcionamiento del instrumento de medida es mediante la elaboración de la curva ROC de la prueba. Las curvas ROC *(Receiver Operating Characteristic),* o en español característica operativa del receptor (COR), tiene sus orígenes en la teoría de la detección de señales (Egan, 1975; Swets, 1996) y permiten analizar la eficacia de los diagnósticos clasificatorios de una prueba a medida que se va variando el punto de corte. Para elaborarla se ubican en el eje de abscisas los valores de la especificidad, en concreto «1 − Especificidad», y en ordenadas, los valores de la sensibilidad. Al ir variando el punto de corte, se obtiene una curva como la de la figura 3.3.

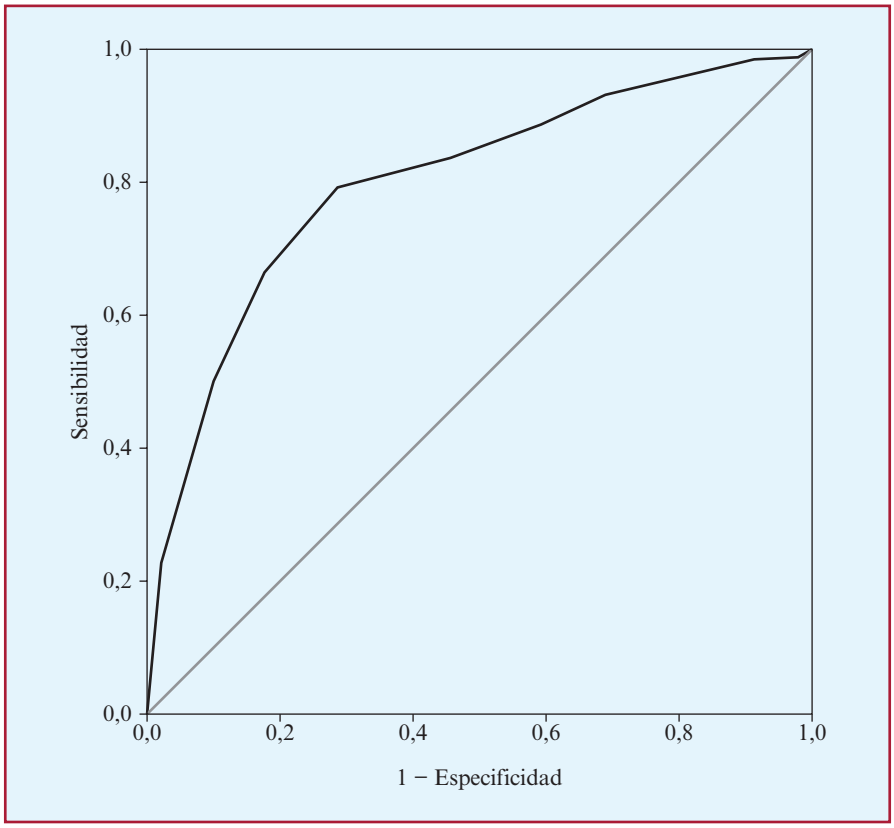


Figura 3.3.—Curva ROC.

Si un test o escala clasificase a las personas al azar en un determinado criterio, la curva ROC coincidiría con la diagonal del cuadrado. A medida que se aleja de la diagonal, indica que mejora la capacidad clasificatoria de la prueba. Una forma de cuantificar este alejamiento es calcular el área que deja la curva ROC por debajo de sí, considerando el cuadrado total como la unidad. Existen numerosos programas para elaborar las curvas ROC y calcular el área que dejan por debajo, que van desde Excel, SAS o SPSS, entre otros. En el caso de la curva del gráfico se utilizó el programa SPSS para calcular el área bajo la curva, que fue de 0,80, con un error típico de 0,02, hallándose el valor poblacional entre los valores 0,757 y 0,844. El valor del área se encuentra lógicamente entre 0 y 1; a modo orientativo se ofrecen a continuación los valores del área bajo la curva para evaluar el funcionamiento del test:

Área	Funcionamiento
0,5-0,6	Malo
0,6-0,7	Flojo
0,7-0,8	Aceptable
0,8-0,9	Bueno
0,9-1,0	Excelente

Estos valores deben tomarse como orientación, pues un mismo valor puede interpretarse de modo diferente dependiendo del contexto y de las decisiones a tomar. El programa también ofrece los datos de la tabla adjunta, que nos indican la sensibilidad y especificidad de la prueba asociadas a cada punto de corte establecido. Esta información es muy valiosa, pues nos permite ubicar el punto de corte en función de los valores a maximizar. Nótese cómo al aumentar la sensibilidad disminuye la especificidad y viceversa. No existe un punto de corte óptimo para toda situación, dependerá de lo que nos interese maximizar en cada caso. Una buena guía aplicada para el uso de las curvas ROC, así como el sofware correspondiente, puede verse en Lasko et al. (2005).

Punto de corte	Sensibilidad	Especificidad
0,00	1,000	0.000
0,50	0,989	0,021
1,50	0,986	0,085
2,50	0,968	0,170
3,50	0,932	0,312
4,50	0,888	0,404
5,50	0,838	0,539
6,50	0,791	0,716
7,50	0,665	0,823
8,50	0,500	0,901
9,50	0,227	0,979
11,00	0,000	1,000

4.4. Selección y clasificación

Las evidencias de validez predictiva de los test, cuya lógica y operativización se analizaron en los apartados precedentes dedicados a la regresión, pueden utilizarse en numerosas áreas aplicadas y profesionales de psicología, amén de su uso e implicaciones de orden teórico. Una de esas áreas de aplicación ha sido clásicamente la selección y clasificación de personal, que posee un estatus propio dentro de las especializaciones profesionales de los psicólogos, con revisiones monográficas y sistemáticas en el *Annual Review* (Dunnette y Borman, 1979; Guion y Gibson, 1988; Hakel, 1986; Tenopyr y Oeltjen, 1982; Zedeck y Cascio, 1984; Hough y Oswald, 2000; Ryan y Ployhart, 2014) y con tecnología altamente especializada y sofisticada (Schmidt y Hunter, 1998).

4.4.1. *Modelos de selección*

Las líneas que siguen de ninguna manera pretenden servir de introducción a esta área de especialización profesional. Únicamente tratan de conectar el modelo de regresión previamente expuesto con su posible uso en alguna de las fases de que constan los complejos procesos de selección y clasificación.

Desde este punto de vista, pueden distinguirse básicamente tres modelos de selección:

— Compensatorio.
— Conjuntivo.
— Disyuntivo.

A los que cabe añadir otros dos de tipo mixto:

— Conjuntivo-compensatorio.
— Disyuntivo-compensatorio.

Veamos en qué consiste cada uno de ellos y cómo funcionan para los dos paradigmas clásicos de la selección:

— Seleccionar un número determinado de personas.
— Seleccionar aquellas personas que superen un cierto nivel de competencia, independientemente de su número.

Modelo compensatorio

Una selección siempre se hace a partir de diversos datos acerca de los aspirantes, algunos de los cuales serán puntuaciones en ciertos test, currículums, entrevistas, etc., en términos generales indicadores varios obtenidos por el psicólogo conectados con el criterio o función a desarrollar por los seleccionados. Ahora bien, si se tienen varias medidas de otros tantos indicadores de competencia, existen muchas posibles formas de combinar esos datos para ordenar a los sujetos según su competencia. Con el modelo compensatorio se lleva a cabo una combinación aditiva de las distintas puntuaciones de las personas, dejando a estas ordenadas según su puntuación global. Claro que hay diversas formas de hacer combinaciones aditivas.

Precisamente la regresión múltiple permite efectuar adecuadamente este tipo de combinación aditiva, escalando a las personas según su puntuación global pronosticada en el criterio (Y'), y asignando a cada predictor el peso pertinente, según se ha expuesto. Una forma mucho más primaria de obtener la puntuación global consistiría en sumar simplemente las puntuaciones de los distintos indicadores, pero esto solo estaría justificado en el caso de que todos ellos tuviesen el mismo peso a la hora de pronosticar el criterio, hecho poco frecuente. Una vez ordenadas las personas por su puntuación global así obtenida, y según el paradigma de selección utilizado, o bien se elige un número determinado, o bien aquellas que superen cierto nivel.

El término «compensatorio» alude a que según este modelo una persona puede compensar su baja competencia en un predictor con una muy buena en otro, dado que lo que se tiene en cuenta es solo el resultado global aditivo. Ahora bien, la compensación no siempre tiene sentido, pues en numerosas situaciones la ausencia de cierta destreza no puede ser compensada con el exceso en otra. Piénsese, por ejemplo, en lo poco afortunado que sería compensar la deficiente coordinación visomotora de un conductor con, digamos, su exhaustivo conocimiento del código de circulación, o la incompetencia técnica en un profesional con sus habilidades sociales.

Modelo conjuntivo

Según este modelo, se seleccionan aquellas personas que superan en todos y cada uno de los predictores un cierto nivel de competencia prefijado, con lo que se evita en gran medida el problema de la compensación de competencias bajas en unos predictores con altas en otros, descartándose aquellos candidatos que no alcancen el citado nivel en cada predictor.

Modelo conjuntivo-compensatorio

Una vez que se han elegido según el modelo conjuntivo las personas que superan cierto nivel en los predictores, se les aplica solo a ellas el modelo compensatorio, quedando así ordenadas según su puntuación global, pudiendo de nuevo o bien elegir cierto número de ellas, o las que superen determinado nivel de puntuación global.

Modelo disyuntivo

Se seleccionan aquellas que superan cierto nivel de competencia en al menos un predictor, es decir, o se supera uno o se supera otro, al menos uno, aunque pueden complejizarse más las exigencias por bloques de predictores. Si a las que cumplan este criterio se les aplica el modelo compensatorio, estaríamos ante el modelo mixto *disyuntivo-compensatorio*.

Los modelos más utilizados en la práctica suelen ser el compensatorio, el conjuntivo y sobre todo el conjuntivo-compensatorio, aunque en determinadas situaciones de selección podría ser más indicado algún otro.

4.4.2. *Utilidad de la selección*

A la hora de evaluar la eficacia de una selección no solo se ha de tener en cuenta la validez de los predictores, tal como se ha venido exponiendo hasta aquí, sino que han de contemplarse, además, aspectos como la razón de selección, la razón de eficacia y la razón de idoneidad.

Se denomina «razón de selección» a la proporción de personas seleccionadas del total de aspirantes, y «razón de eficacia» a la proporción de seleccionados que efectivamente tienen éxito posterior en el criterio.

Por ejemplo, si de 200 aspirantes a tornero seleccionamos 50 y de estos solo 20 resultaron buenos torneros, ¿cuánto valen la razón de selección y la razón de eficacia?

— Razón de selección:

$$\frac{50}{200} = 0{,}25$$

— Razón de eficacia:

$$\frac{20}{50} = 0{,}40$$

Se entiende por razón de idoneidad la proporción de aspirantes cualificados para tener éxito en el criterio. Lógicamente este dato no se conoce directamente, por lo que solo cabe hacer algunas estimaciones.

Taylor y Russell (1939) elaboraron unas tablas (véase tabla 3.2), hoy clásicas, que para un valor estimado de la razón de idoneidad, y conocidas la validez y la razón de selección, permiten estimar cuál sería la razón de eficacia o probabilidad de que un sujeto seleccionado bajo esas circunstancias tenga éxito, lo cual constituye un indicador de la utilidad de la selección, entendiendo por utilidad aquello que la selección mejora la razón de eficacia respecto al azar o a otro tipo de selección.

TABLA 3.2

Tabla de Taylor-Russell para una razón de idoneidad de 0,50

r	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
0,00	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50
0,05	0,54	0,54	0,53	0,52	0,52	0,52	0,51	0,51	0,51	0,50	0,50
0,10	0,58	0,57	0,56	0,55	0,54	0,53	0,53	0,52	0,51	0,51	0,50
0,15	0,63	0,61	0,58	0,57	0,56	0,55	0,54	0,53	0,52	0,51	0,51
0,20	0,67	0,64	0,61	0,59	0,58	0,56	0,55	0,54	0,53	0,52	0,51
0,25	0,70	0,67	0,64	0,62	0,60	0,58	0,56	0,55	0,54	0,52	0,51
0,30	0,74	0,71	0,67	0,64	0,62	0,60	0,58	0,56	0,54	0,52	0,51
0,35	0,78	0,74	0,70	0,66	0,64	0,61	0,59	0,57	0,55	0,53	0,51
0,40	0,82	0,78	0,73	0,69	0,66	0,63	0,61	0,58	0,56	0,53	0,52
0,45	0,85	0,81	0,75	0,71	0,68	0,65	0,62	0,59	0,56	0,53	0,52
0,50	0,88	0,84	0,78	0,74	0,70	0,67	0,63	0,60	0,57	0,54	0,52
0,55	0,91	0,87	0,81	0,76	0,72	0,69	0,65	0,61	0,58	0,54	0,52
0,60	0,94	0,90	0,84	0,79	0,75	0,70	0,66	0,62	0,59	0,54	0,52
0,65	0,96	0,92	0,87	0,82	0,77	0,73	0,68	0,64	0,59	0,55	0,52
0,70	0,98	0,95	0,90	0,85	0,80	0,75	0,70	0,65	0,60	0,55	0,53
0,75	0,99	0,97	0,92	0,87	0,82	0,77	0,72	0,66	0,61	0,55	0,53
0,80	1,00	0,99	0,95	0,90	0,85	0,80	0,73	0,67	0,61	0,55	0,53
0,85	1,00	0,99	0,97	0,94	0,88	0,82	0,76	0,69	0,62	0,55	0,53
0,90	1,00	1,00	0,99	0,97	0,92	0,86	0,78	0,70	0,62	0,56	0,53
0,95	1,00	1,00	1,00	0,99	0,96	0,90	0,81	0,71	0,63	0,56	0,53
1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,83	0,71	0,63	0,56	0,53

Un análisis de la utilidad de la selección solo puede llevarse a cabo de un modo riguroso aplicando la teoría estadística de la decisión. Exposiciones introductorias dirigidas a psicólogos pueden consultarse en Wiggins (1973), Coombs, Dawes y Tversky (1981), o más extensamente en Cronbach y Glesser (1965).

Obsérvese en la tabla de Taylor-Russell correspondiente a una razón de idoneidad de 0,50 cómo incluso con una validez baja, cuando la razón de selección es reducida, la probabilidad de éxito en el criterio (razón de eficacia) es alta.

Por ejemplo, con una validez ínfima de 0,25, si la razón de selección es 0,05, la probabilidad de éxito es de 0,70; luego se ha ganado un 20% respecto a si la selección se hubiese hecho al azar, en cuyo caso la probabilidad de éxito hubiese sido de 0,50, ya que para esta tabla la proporción de aspirantes idóneos (razón de idoneidad) es 0,50.

En la tabla 3.2, la columna de números debajo de *r* corresponde a la validez del test, la fila superior de números a la razón de selección, y los números interiores son las razones de eficacia correspondientes.

Taylor y Russell ofrecen tablas como la representada aquí para valores de la razón de idoneidad de 0,05, 0,10, 0,20, 0,30, 0,40, 0,50 (reproducida aquí), 0,60, 0,70, 0,80, 0,90 y 0,95. No conviene olvidar que dichas tablas asumen una distribución test-criterio bivariada normal, lo cual es discutible que se cumpla siempre empíricamente. En cualquier caso, las tablas constituyen un marco de referencia orientador y sencillo en el proceso de selección, mucho más complejo que la mera aplicación de test y la predicción automática del criterio.

Una idea de la eficacia de un test para pronosticar el criterio en el proceso de selección puede obtenerse también mediante una tabla de doble entrada donde figuren los pronósticos realizados a partir del test y los resultados que realmente se han dado posteriormente. Véase, por ejemplo, la tabla 3.3.

Nótese que de las 30 personas seleccionadas por el test solo 10 tuvieron éxito en el criterio, y de las 70 no seleccionadas lo tuvieron 30. La tasa de aciertos del test es solo del 50%: [(10 + 40)/100 = 0,50]. Manipulando el punto de corte en el test, se puede variar la tasa de aciertos, buscándose aquel punto de corte que maximiza la tasa.

TABLA 3.3

		Resultados reales en el criterio		
		Éxitos	**Fracasos**	
Pronósticos del test	Seleccionados	10	20	30
	No seleccionados	30	40	70
		40	60	100

Ha de tenerse en cuenta, además, que por razones inherentes a la propia naturaleza de la selección concreta que se esté haciendo, no siempre tienen la misma importancia para el psicólogo los dos tipos de errores posibles, a saber, aquellos seleccionados que fracasan (en la tabla, 20) y los no seleccionados que tienen éxito (en la tabla, 30). Según se considere más grave un error u otro, puede variar notablemente la estrategia selectiva.

Como ya se ha señalado, la teoría de la decisión constituye el marco teórico imprescindible que permite considerar toda la casuística apuntada aquí de un modo sistemático y coherente.

Uso del modelo de regresión

Aunque el coeficiente de validez de las pruebas utilizadas en la selección sea elevado, ello no garantiza que las personas seleccionadas tengan éxito seguro en el criterio, simplemente aumenta (más o menos) la probabilidad de que ocurra. El modelo de regresión permite estimar estas probabilidades de éxito si la situación se ajusta a las condiciones impuestas por el modelo. Veamos cómo se procede en un caso sencillo.

Ejemplo

Supóngase un test (X) cuya recta de regresión para predecir el criterio (Y) viene dada por $Y' = X/2 + 4$, siendo 10 la desviación típica del criterio ($S_y = 10$) y 0,80 el coeficiente de validez del test. Si se considera como criterio de éxito en el criterio superar los 9 puntos en este, ¿qué probabilidad de éxito tendrán

las personas que hayan sacado en el test 16 puntos? (véase figura 3.4).

$$Y' = \frac{X}{2} + 4$$

Para un valor de $X = 16$:

$$Y' = \frac{16}{2} + 4 = 12$$

La desviación típica de los errores de estimación o error típico de estimación ($S_{y \cdot x}$), que se asumen distribuidos según la curva normal e iguales (homoscedasticidad), viene dada por:

$$S_{y \cdot x} = S_y \sqrt{(1 - r_{xy}^2)} = 10\sqrt{(1 - 0{,}80)^2} = 6$$

La puntuación típica (Z_c) correspondiente al valor del criterio (9) será:

$$Z_c = \frac{9 - 12}{6} = -0{,}50$$

En las tablas de la curva normal por encima de −0,50 queda una proporción de 0,6915; luego la persona que sacó 16 puntos sí ha sido seleccionado, ya que se le pronostica una puntuación en el criterio de 12, por encima de los 9 exigidos, pero no es seguro que tenga éxito. En concreto, se le asigna una probabilidad de éxito de 0,6915 y, por tanto, una probabilidad de fracaso de 0,3085.

4.4.3. *Clasificación*

El problema de la clasificación en psicología podría considerarse en cierto modo un caso particular de la predicción en el que el objetivo es asignar las personas a determinadas categorías, ya sean cuadros diagnósticos, profesiones, etc. Se trataría, en suma, de predecir qué categoría es pertinente en cada caso, según las variables consideradas, y maximizar la probabilidad de categorización correcta. Se asume, naturalmente, que esas categorías han sido validadas empíricamente y que incluir en una u otra tiene implicaciones de interés, y no constituye un mero ejercicio de etiquetado. Las categorías, clasificaciones o cuadros han de ser relevantes para algo; de lo contrario, cambiar el nombre propio de la

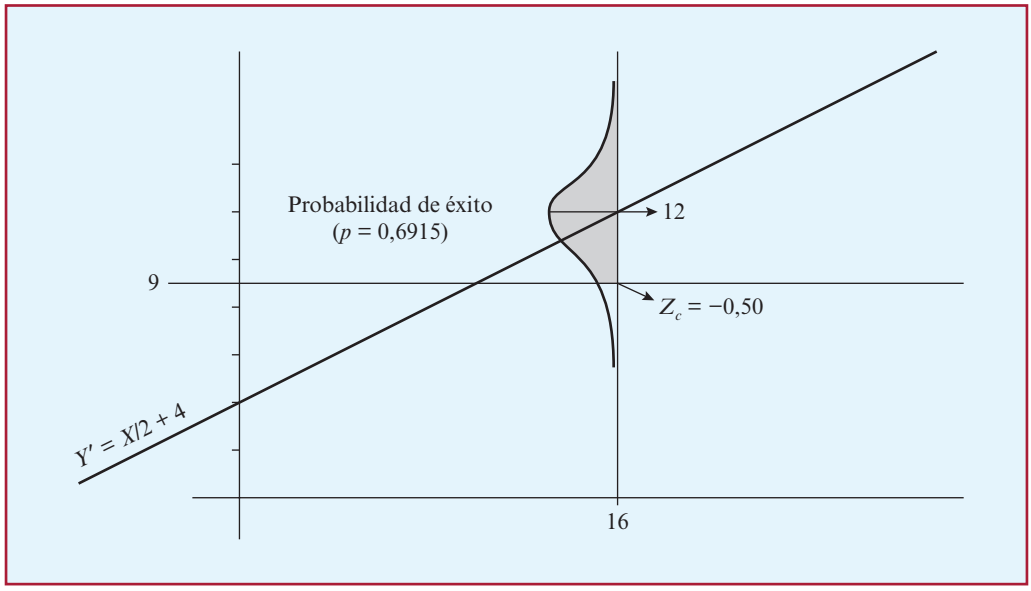


Figura 3.4.

persona por el de una rumbosa categoría no tiene ningún sentido. No hay duda de que en la actividad profesional de los psicólogos áreas como la orientación profesional/escolar, la selección o el diagnóstico tienen mucho que ver con el problema general de hacer corresponder lo exigido por determinada situación con las características de las personas.

Todo psicólogo que se dedique a estos menesteres, aparte del bagaje sustantivo del área correspondiente, es obligado que se ayude de algunas técnicas estadísticas multivariadas que pueden mejorar significativamente sus decisiones. En concreto, y además de la regresión, ya comentada, el análisis discriminante y el análisis de *cluster* proporcionan mejoras sustantivas al sentido común y la experiencia.

Lejos de tratar de exponer estas técnicas aquí, se citará brevemente qué tipo de problema permiten resolver y la bibliografía donde se pueden consultar.

El análisis discriminante tiene bastante semejanza conceptual con la regresión múltiple. Se miden determinadas variables predictoras a partir de las cuales se elabora una función discriminante, o más, en las que los pesos de las predictoras están elaborados de tal guisa que maximicen la asignación correcta de las personas a ciertas categorías previamente establecidas. Lo que en la regresión múltiple era el criterio a predecir aquí son las categorías dentro de las cuales se desea clasificar a las personas. En suma, el análisis discriminante permite clasificar a las personas en categorías, a partir de sus puntuaciones en determinadas variables predictoras que se ponderan adecuadamente para maximizar los aciertos en la clasificación. Esta técnica, implementada, entre otros, en el programa informático SPSS, permite evaluar no solo el porcentaje de clasificaciones correctas que se hacen, sino la relevancia relativa (peso) de las variables predictoras a la hora de la predicción. Una introducción sencilla pude consultarse en Pedhazur (1982), y descripciones más detalladas, en Klecka (1980) o Tatsuoka (1970); para aplicaciones véase Huberty (1975).

El análisis de *cluster* permite formar conglomerados o grupos de personas, u otros entes, semejantes entre sí. Para ello hay que establecer previamente alguna medida o indicador del grado de afinidad o asociación entre las personas, que será el dato básico que permitirá establecer los *clusters,* asignándose el mismo *cluster* a las personas más parecidas entre sí en función de la medida de similitud utilizada. Esta medida dependerá de los objetivos de la investigación que se lleve a cabo; por ejemplo, podría muy bien ser la correlación entre los perfiles psicológicos de las personas, en cuyo caso los *clusters* resultantes estarían formados por aquellas personas cuyos perfiles fuesen más parecidos. La técnica también permite hacer ese mismo agrupamiento, en vez de con las personas, con las variables, en cuyo caso su semejanza de objetivos con el análisis factorial es clara: lo que allí eran factores aquí serían *clusters.* Véase para exposiciones detalladas y no excesivamente técnicas Anderberg (1973), Everitt (1974), Hartigan (1975), Lorr (1983), Milligan y Cooper (1987) o Spath (1980).

Aún cabría citar otras técnicas multivariadas que son de gran ayuda en esta área, tales como la correlación canónica o los modelos de ecuaciones estructurales, amén del análisis factorial, claro está.

Coeficiente *kappa*

Cuando se llevan a cabo clasificaciones hechas por distintos métodos o distintos clasificadores humanos, siempre aparece el problema de determinar en qué grado hay acuerdo entre dichas clasificaciones; en definitiva, el problema de la fiabilidad de las clasificaciones. Caso típico puede ser el análisis de los acuerdos-desacuerdos entre los diagnósticos psicológicos hechos por diferentes profesionales, clasificaciones de alumnos por distintos profesores, pacientes, etc.

La estadística proporciona numerosos índices para objetivar el grado de asociación entre este tipo de variables (Ato, 1991; Haberman, 1974, 1978; Smith, 1976; Everitt, 1977; Fienberg, 1977), pero no se puede dejar de comentar el popular coeficiente *kappa* de Cohen (1960).

Como ya se ha visto al tratar la fiabilidad de los test referidos al criterio, la fórmula del coeficiente *kappa* viene dada por:

$$K = \frac{F_c - F_a}{N - F_a} \qquad [3.40]$$

donde

F_c: Número de casos (frecuencia) en los que ambos clasificadores coinciden.

F_a: Número de casos (frecuencia) en que cabe esperar por mero azar que los clasificadores coincidan.
N: Número total de casos.

Si en vez de en frecuencias se expresa en términos de proporciones, el coeficiente vendrá dado por:

$$K = \frac{P_c - P_a}{1 - P_a} \qquad [3.41]$$

donde P_c y P_a son ahora proporciones de coincidencia y azar, respectivamente, en vez de frecuencias.

EJEMPLO

Veamos el ejemplo con el que Cohen (1960) presentó su afamado coeficiente. Es el caso de dos jueces que han de clasificar 200 sujetos en tres categorías. Los datos aparecen en la tabla 3.4; ¿cuál es el grado de acuerdo entre los juicios de ambos jueces?

Las frecuencias entre paréntesis son aquellas que cabría esperar por mero azar:

$$C_{11} = \frac{(120)(100)}{200} = 60$$

$$C_{22} = \frac{(60)(60)}{200} = 18$$

$$C_{33} = \frac{(20)(40)}{200} = 4$$

Por tanto:

$$F_c = 88 + 40 + 12 = 140$$

$$F_a = 60 + 18 + 4 = 82$$

Aplicando la fórmula propuesta:

$$K = \frac{140 - 82}{200 - 82} = 0{,}492$$

El valor máximo del coeficiente *kappa* es +1, pero el valor mínimo no es −1, sino que depende de las frecuencias marginales.

TABLA 3.4

		Juez *A*			
		Categoría 1	**Categoría 2**	**Categoría 3**	
Juez *B*	**Categoría 1**	88 (60)	14	18	120
	Categoría 2	10	40 (18)	10	60
	Categoría 3	2	6	12 (4)	20
		100	60	40	200

EJERCICIOS

1. Utilizando una muestra de 1.000 personas, un psicólogo encontró que el coeficiente de fiabilidad de una escala de paranoidismo era 0,75 y el coeficiente de fiabilidad del criterio era de 0,80. La correlación entre las puntuaciones obtenidas por los sujetos en la escala y las obtenidas en el criterio resultó ser de 0,70.

1. Calcular el coeficiente de validez de la escala.
2. Calcular el porcentaje de varianza del criterio pronosticable a partir de la escala.
3. ¿Cuánto valdría el coeficiente de validez de la escala si esta careciese de errores de medida?

4. ¿Cuál sería el coeficiente de validez de la escala si la fiabilidad del criterio fuese perfecta?
5. ¿Cuál sería el coeficiente de validez de la escala si se eliminasen todos los errores de medida tanto de esta como del criterio?
6. Si se lograse por algún medio elevar la fiabilidad de la escala hasta 0,85, ¿cuánto valdría su coeficiente de validez?
7. ¿Cuál sería el coeficiente de validez de la escala si la fiabilidad del criterio se mejorase hasta alcanzar el valor de 0,95?
8. En el caso de llevar a cabo conjuntamente las dos mejoras anteriores de la fiabilidad de la escala y del criterio, ¿cuál sería el coeficiente de validez de la escala?

2. La correlación entre las puntuaciones de 500 sujetos en un test de inteligencia espacial que consta de 25 ítems y las obtenidas por ellos en una tarea empírica considerada como criterio fue de 0,60. La correlación entre las citadas puntuaciones de los sujetos en el test y las puntuaciones que estos mismos sujetos obtuvieron de nuevo en el test al aplicárselo tres días después fue de 0,65.

1. ¿Cuál sería el coeficiente de validez del test si se le añadiesen otros 20 ítems paralelos?
2. ¿Cuántos ítems habrá que añadir a los 25 originales si se desea obtener un coeficiente de validez de 0,70?
3. ¿Cuántos de los 25 ítems originales se podrían suprimir si nos conformásemos con una validez de 0,50?
4. ¿Sería posible obtener un coeficiente de validez de 0,95 a base de aumentar la longitud del test? Razone adecuadamente.
5. Represente gráficamente el incremento del coeficiente de validez (ordenadas) al aumentar el número de ítems de 10 en 10 (abscisas) hasta 100 ítems. Comente el resultado.

3. Demuestre que la correlación entre las puntuaciones en el criterio (y) y las pronosticadas en él (y') es igual a la correlación entre las puntuaciones del test (x) y del criterio (y). Es decir, demuestre que $r_{yy'} = r_{xy}$.

4. Demuestre que $r_{xy'} = 1$.

5. Demuestre que $r_{(y - y')x} = 0$.

6. La media de una muestra de 1.000 sujetos en un test de rapidez perceptiva fue de 20 puntos, y la desviación típica, de 4, con un 36% de varianza asociada entre test y criterio. Por su parte, la media del criterio es 30, y la varianza, 25.

1. ¿Qué puntuación se estima que obtendrán en el criterio aquellos sujetos que obtuvieron 15 puntos en el test? Nivel de confianza del 99%.
2. De los sujetos anteriores que obtuvieron 15 puntos en el test, ¿qué proporción de ellos cabe esperar que obtengan en el criterio puntuaciones iguales o mayores que la media de este?
3. Todo igual, ¿a qué nivel de confianza habría que trabajar para que el error máximo admisible no excediese de 6?
4. Calcule el valor de la pendiente y de la ordenada en el origen de la recta de regresión del criterio sobre el test.

7. La varianza de las puntuaciones globales (G) obtenidas al sumar las obtenidas por una muestra de sujetos en cierto test (X) y las obtenidas en el criterio (Y) vale 61 ($G = X + Y$, $S_G^2 = 61$). La varianza de la diferencia es 21 ($D = X - Y$, $S_D^2 = 21$). La desviación típica del test es el 80% de la del criterio.

1. Calcular la varianza del test y del criterio.
2. Calcular el coeficiente de validez del test.

8. La desviación típica de las puntuaciones de los 100 alumnos que mejor calificación obtuvieron en una práctica de laboratorio sobre aprendizaje animal fue de 15, mientras que la varianza de las calificaciones de todos los que realizaron la práctica fue de 400. Solo a los 100 mejores alumnos se les aplicó una escala de extraversión-introversión, obteniéndose una varianza de las puntuaciones de 25. Siendo X las puntuaciones de los sujetos en la escala e Y sus puntuaciones en la práctica, la varianza de la suma de ambas puntuaciones resultó ser 340.

1. Calcular el coeficiente de validez de la escala para predecir el rendimiento de los alumnos en las prácticas de laboratorio.

9. Para la prueba de admisión en la escuela de arquitectura se les aplica a los aspirantes un test de aptitudes espaciales. La desviación típica de las puntuaciones en el test de los aspirantes fue de 25 y la varianza de las puntuaciones de los admitidos (solo el 10%) fue de 2. El coeficiente de valor predictivo en el grupo de admitidos es 0,5641, tomando como criterio las calificaciones de la carrera. También se aplica a los admitidos otro test de razonamiento espacial que está en fase experimental para su posible uso en la selección, obteniéndose que el cociente entre el error típico de estimación de este nuevo test y la desviación típica del criterio fue 0,50, con una correlación entre ambos test de 0,40.

1. A la vista de estos datos, ¿puede afirmarse que el nuevo test en fase experimental mejora los pronósticos del que ya se viene utilizando? NC del 95%.

10. A los 25.000 médicos aspirantes al MIR se les aplica un test de conocimientos como prueba de selección. La varianza de sus puntuaciones en esta prueba fue de 324. La desviación típica de los admitidos es 1/9 de la desviación típica obtenida para el grupo de aspirantes, de los que, por cierto, solo se admitió a un 20%. Terminado el período de MIR, los admitidos fueron valorados por sus profesores según su eficacia en el trabajo (criterio), resultando que solo el 9% de la varianza de la eficacia era pronosticable a partir del test de selección. Ante este porcentaje tan bajo, se confecciona otro test que sirva de alternativa en la selección, y tras aplicarlo a este grupo de médicos se encuentra una correlación con el criterio anterior de 0,50, lo cual resulta a priori esperanzador, dado lo restringido del grupo en el que se calculó. La correlación entre las puntuaciones de los sujetos en ambos test resultó ciertamente baja: 0,15.

1. ¿Puede afirmarse que el nuevo test es preferible al anterior para efectuar la selección? NC del 95%.

11. Para ilustrar el proceso de validación propuesto, Campbell y Fiske (1959) presentan la matriz multirrasgo-multimétodo que se reproduce aquí, en la que se han medido tres rasgos (*A*, *B*, C) por tres métodos diferentes. (Las correlaciones aparecen multiplicadas por 100.)

		Método 1			**Método 2**			**Método 3**		
		A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Método 1	A_1	89								
	B_1	51	89							
	C_1	38	37	76						
Método 2	A_2	57	22	09	93					
	B_2	22	57	10	68	94				
	C_2	11	11	46	59	58	84			
Método 3	A_3	56	22	11	67	42	33	94		
	B_3	23	58	12	43	66	34	67	92	
	C_3	11	11	45	34	32	58	58	60	85

1. ¿Qué método genera coeficientes de fiabilidad ligeramente inferiores?
2. ¿Entre qué valores se encuentran las correlaciones indicadoras de validez convergente?
3. ¿Existe una buena validez discriminante? Razone adecuadamente.
4. ¿Cuál de los tres métodos maximiza la correlación entre los rasgos *A* y *C*?

12. En la matriz adjunta se presentan los datos (Green, 1976) obtenidos por 12 sujetos en un criterio (Y) y en dos variables predictoras (X_1, X_2).

Y	X_1	X_2
1	1	1
0	2	1
1	2	2
4	3	2
3	5	4
2	5	6
5	6	5
6	7	4
9	10	8
13	11	7
15	11	9
16	12	10

1. Obtenga las ecuaciones de regresión en puntuaciones directas, diferenciales y típicas para pronosticar Y a partir de X_1 y X_2.
2. Calcule la correlación múltiple.
3. ¿Qué porcentaje de la varianza del criterio viene explicada por las variables predictoras?
4. Calcule el error típico de estimación.
5. Calcule el valor de la varianza asociada y el de la no asociada.
6. ¿Cuál sería la correlación entre el criterio (Y) y la variable predictora X_1 si se controlase el efecto de X_2 sobre ambas?

13. En esta matriz aparecen las puntuaciones de cinco sujetos en dos test paralelos, A y A', así como las puntuaciones obtenidas por los sujetos en el criterio.

	Test A				Test A'				Puntuaciones criterio
	Ítems				Ítems				
	1	2	3	4	1	2	3	4	
A	1	0	0	0	0	0	0	0	2
B	1	1	0	0	1	1	0	0	4
C	1	1	1	0	1	1	1	1	6
D	1	1	1	1	1	1	1	0	8
E	0	0	0	0	1	0	0	0	0

1. Calcule el coeficiente de validez del test A y comente el resultado.
2. Al NC del 99%, ¿entre qué puntuaciones se encontrará en el criterio la de aquellos sujetos que obtuviesen 2 puntos en el test A?
3. Elabore la ecuación de regresión en puntuaciones típicas que permite pronosticar el criterio a partir de ambos test. Comente las diferencias, si hubiese, entre los valores de los pesos *beta* para cada test.
4. Calcule el porcentaje de varianza del criterio pronosticable a partir de ambos test.

14. Un psicólogo escolar está investigando la capacidad predictora de tres pruebas psicológicas, inteligencia (IG), motivación de logro (ML) y ajuste emocional (AE), para predecir el rendimiento académico (RA) de los estudiantes de su centro. Las correlaciones entre las puntuaciones de las cuatro variables, obtenidas en una muestra de 500 estudiantes, aparecen en la matriz adjunta.

	RA	IG	ML	AE
RA	1,00			
IG	0,60	1,00		
ML	0,50	0,00	1,00	
AE	0,25	0,00	0,00	1,00

1. Elabore la ecuación de regresión en puntuaciones típicas para pronosticar el RA a partir de los tres test. Especifique todos los pasos seguidos.
2. ¿Qué porcentaje de la varianza del RA viene explicada por las tres variables predictoras?
3. Al nivel de confianza del 95%, ¿qué puntuación se le pronosticará en RA a un sujeto que obtuvo en las variables predictoras las siguientes puntuaciones típicas: $IG = 0{,}30$, $ML = 0{,}40$, $AE = 0{,}20$.
4. ¿Puede afirmarse que la motivación de logro tiene un efecto modulador sobre las relaciones entre la IG y el RA?

15. En una muestra de 500 sujetos, la ecuación de regresión en puntuaciones típicas para pronosticar cierto criterio (Y) a partir de cuatro test (X_1, X_2, X_3, X_4) resultó ser:

$$Z_{y'} = 0{,}8Z_{x_1} + 0{,}7Z_{x_2} + 0{,}5Z_{x_3} + 0{,}2Z_{x_4}$$

Por su parte, las correlaciones de estas cuatro variables con el criterio fueron:

$$r_{x_1y} = 0{,}4;\ r_{x_2y} = 0{,}3;\ r_{x_3y} = 0{,}25;\ r_{x_4y} = 0{,}1$$

1. Calcular la correlación múltiple.
2. Al nivel de confianza del 95%, ¿puede afirmarse que la correlación múltiple es estadísticamente significativa?

16. A una reciente oferta de trabajo publicada en *El País* para especialistas en técnicas didácticas se presentaron 400 licenciados universitarios, de los que solo fueron admitidos los 20 que obtuvieron mejores puntuaciones en un test selector. Las puntuaciones de los aspirantes en ese test se distribuyeron según la curva normal, con una media de 60 y una desviación típica de 4. El test resultó tener un coeficiente de validez de 0,80 respecto a un criterio con una varianza de 36 y una media de 100.

1. ¿Cuál fue la razón de selección?
2. ¿Cuál es la puntuación directa que como mínimo deben haber obtenido en el test los seleccionados?
3. Para que un aspirante haya sido admitido, ¿qué puntuación directa mínima se le debe pronosticar en el criterio?
4. ¿Qué puntuación directa obtendrá en el criterio un sujeto que solo fue superado en el test por 10 de sus compañeros?
5. ¿Cuál es la probabilidad de que fracase en su cometido (criterio) un sujeto que obtuvo en el test una puntuación directa de 70 puntos y que, por tanto, fue seleccionado?

17. Las relaciones entre la rapidez para procesar información (RPI) y la inteligencia verbal han sido ampliamente investigadas, entre otros, por Hunt y colaboradores. En una prueba de RPI, una muestra de 1.000 sujetos obtuvo una media de 300 milisegundos y una desviación típica de 20, siendo el coeficiente de fiabilidad de la prueba de 0,60 (la prueba de RPI constaba de 50 ítems). Aplicado a la misma muestra de sujetos el test verbal SAT, se obtuvo una media de 40 y una desviación típica de 5, siendo el coeficiente de fiabilidad del SAT de 0,70. El porcentaje de varianza asociada entre la prueba de RPI y el SAT (criterio) fue del 64%.

1. Exprese mediante algún índice numérico la capacidad de la RPI para predecir las puntuaciones del SAT.
2. ¿Cuál sería la validez de la prueba de RPI si las mediciones del SAT careciesen totalmente de errores de medida?
3. Se comprobó que los 50 ítems de la prueba de RPI eran pocos, por lo que se añadieron otros 25 similares (paralelos) a los que ya poseía. ¿Cuál será la validez de la prueba una vez alargada?
4. Elabore la ecuación de regresión que permite pronosticar las puntuaciones del SAT a partir de las obtenidas en la prueba de RPI.
5. Un sujeto que obtuvo en la prueba de RPI una puntuación típica de 1,75 ¿qué puntuación directa obtendrá en el SAT? NC del 95%.

18. Vernon (1983) investigó las posibilidades de pronosticar la inteligencia (IG) a partir de los siguientes procesos básicos: tiempo de reacción (TR), tiempo de inspección (TI), memoria a corto plazo (MCP) y memoria a largo plazo (MLP). La matriz de correlaciones obtenidas (datos no reales) entre las cinco variables citadas aparecen a continuación:

1. Elabore la ecuación de regresión que permita pronosticar la IG a partir de los cuatro procesos básicos.
2. Calcule la correlación múltiple entre las cuatro variables predictoras y la IG.
3. ¿Qué porcentaje de varianza de la IG es pronosticable a partir de los cuatro procesos básicos?

	IG	TR	TI	MCP	MLP
IG	1,00				
TR	0,50	1			
TI	0,30	0	1		
MCP	0,20	0	0	1	
MLP	0,25	0	0	0	1

4. ¿Qué puntuación se estima que obtendrán en inteligencia aquellos sujetos que hubiesen obtenido en las variables predictoras las siguientes puntuaciones típicas: $Z_{TR} = 0{,}45$,

$Z_{TI} = 0{,}35$, $Z_{MCP} = 0{,}15$, $Z_{MLP} = 0{,}20$? NC del 95%.

5. Si deseásemos una ecuación de regresión con solo tres variables predictoras y, en consecuencia, eliminásemos una por el método *backward,* ¿qué variable eliminaríamos? Razone adecuadamente.

19. Una empresa que necesita ampliar su plantilla aplica un test de inteligencia a los 50 aspirantes para los 10 puestos a cubrir. Las puntuaciones en el test se distribuyeron según la curva normal con una media de 100 y una desviación típica de 15. El 64% de la varianza del criterio es pronosticable a partir del test.

1. Si el punto crítico de éxito en el criterio se fija dos desviaciones típicas por encima de la media, ¿qué probabilidad de éxito tendrán los sujetos que obtengan en el test una puntuación diferencial de 20 puntos?

20. Un grupo de aspirantes a controladores aéreos obtienen en un test de selección *A* una varianza de 25, y los admitidos, de 4. La correlación del test con el criterio en el grupo de seleccionados fue de 0,60. El equipo de psicólogos del departamento de selección está estudiando la mejora de esta, y para ello investiga la adecuación selectiva de otro test *B.* Este nuevo test experimental se aplicó a los seleccionados con el *A,* obteniéndose una correlación de 0,65 con el criterio y de 0,50 con el test selector *A.*

1. Dado que este nuevo test *B* que se está estudiando tiene una correlación de 0,65 con el criterio, mientras que el *A* usado para seleccionar solo la tiene de 0,60, un estudiante de psicología que realiza prácticas en el departamento de selección opina que el test *B,* obviamente, es mejor que el *A* para hacer la selección. ¿Comparte usted su opinión?

21. Se aplicaron tres test de aptitudes (X_1, X_2, X_3) y una prueba de rendimiento (Y), considerada como criterio, a una muestra de 500 sujetos. La matriz de correlaciones entre estas variables, así como sus varianzas, aparecen a continuación.

1. Calcule el error típico de medida de cada test, así como del criterio.
2. Calcule el error típico de estimación de cada test.
3. Elabore las ecuaciones de regresión de cada test en puntuaciones diferenciales.

Matriz de correlaciones				
	y	X_1	X_2	X_3
Y	0,95			
X_1	0,60	0,90		
X_2	0,40	0,00	0,80	
X_3	0,30	0,00	0,00	0,85

Varianzas
16
9
4
1

4. Elabore la ecuación de regresión múltiple que permite pronosticar el criterio a partir de los tres test considerados conjuntamente. Detalle el proceso de obtención.
5. ¿Qué porcentaje de la varianza del criterio viene explicado por los tres test tomados conjuntamente?
6. Calcule la matriz de varianzas-covarianzas de las cuatro variables (test y criterio).

22. En la tabla adjunta aparecen las puntuaciones obtenidas por una muestra de 15 personas en una escala de psicoticismo de 20 puntos aplicada por el médico de familia de un centro de atención primaria. Esas mismas personas fueron diagnosticadas por un equipo de psicólogos y psiquiatras en dos categorías: psicóticos (PS) y no psicóticos (NP). Los resultados aparecen en la tabla adjunta:

Personas	Escala	Diagnóstico
A	13	*PS*
B	15	*PS*
C	15	*PS*
D	17	*NP*
E	16	*PS*
F	18	*NP*
G	13	*NP*
H	16	*PS*
I	12	*NP*
J	10	*NP*
K	17	*PS*
L	9	*NP*
M	18	*PS*
N	14	*NP*
Ñ	19	*PS*

1. Si se asume que todos los errores son igualmente relevantes:

   1.1. ¿Dónde habría que establecer el punto de corte para minimizar los errores totales de clasificación cometidos al usar la escala para predecir psicoticismo?

   1.2. Establecido el punto de corte del apartado anterior, calcule la proporción total de clasificaciones correctas, sensibilidad, especificidad y coeficiente *kappa* de la escala.

2. Imagínese que los profesionales del campo, psicólogos y psiquiatras, consideran cuatro veces más grave no detectar una persona psicótica que sí lo es que considerar psicótica a una que realmente no lo es. Bajo este supuesto:

   2.1. ¿Dónde se establecería el punto de corte de la escala que minimiza los errores totales?

   2.2. Calcule la proporción total de clasificaciones correctas, sensibilidad, especificidad y coeficiente *kappa* de la escala para el punto de corte del apartado anterior.

3. Asignando determinados pesos a los falsos positivos y a los falsos negativos, se obtienen para los puntos de corte 12 y 16 valores totales de los errores de 15 y 12 respectivamente.

   3.1. ¿Qué pesos se han asignado a los falsos positivos y a los falsos negativos?

4. Describa dos situaciones en el campo de la psicología del trabajo en las que los errores falsos positivos sean mucho más graves que los falsos negativos. Describa otras dos en las que ocurra justo lo contrario.

SOLUCIONES

1.1. 0,70
2. 49%
3. 0,808
4. 0,783
5. 0,90
6. 0,745
7. 0,763
8. 0,812

2.1. 0,653
2. 79
3. 13
4. No. Máximo: 0,744

6.1. $15{,}93 \leqslant Y \leqslant 36{,}57$
2. 0,1736
3. 86,64%
4. $b = 0{,}75$; $a = 15$

7.1. $s_x^2 = 16$; $s_y^2 = 25$
2. 0,50

8.1. 0,71

9.1. No. $R_{ZY} = R_{XY} = 0{,}99$

10.1. No. $R_{XY} = 0{,}94$; $R_{ZY} = 0{,}86$

11.1. 1
2. $0{,}45 \leqslant r \leqslant 0{,}67$
3. No
4. 2

12.1. $Y' = 1{,}55X_1 - 0{,}24X_2 - 2{,}20$
$y' = 1{,}55x_1 - 0{,}24x_2$
$Z_{y'} = 1{,}07Z_1 - 0{,}13Z_2$
2. 0,95
3. 90%
4. 1,71
5. 26,53; 2,95
6. 0,73

13.1. 1
2. $Y' = Y = 4$
3. $Z_{y'} = 1{,}00Z_A + 0{,}00Z_{A'}$
4. 100%

14.1. $Z_{y'} = 0{,}60Z_{IG} + 0{,}50Z_{ML} + 0{,}25Z_{AE}$
2. 67%
3. $-0{,}69 \leqslant Z_y \leqslant 1{,}55$
4. Sí. $r_{xy.z} = 0{,}69$

15.1. 0,82

2. Sí: $F = 257$, $p < 0{,}05$

16.1. 0,05
2. 66,56
3. 107,872
4. 109,408
5. 0,1251

17.1. $r_{xy} = 0{,}80$
2. 0,95
3. 0,86
4. $Y' = 0{,}2X - 20$
5. $41{,}12 \leqslant Y \leqslant 52{,}88$

18.1. $Z_{y'} = 0{,}50Z_1 + 0{,}30Z_2 + 0{,}20Z_3 + 0{,}25Z_4$
2. 0,665
3. 44,25%
4. $-1{,}0535 \leqslant IG \leqslant 1{,}8735$
5. MCP

19.1. 0,0594

20.1. No. $R_{ZY} = 0{,}86$; $R_{XY} = 0{,}88$

21.1. $S_e(x_1) = 0{,}95$; $S_e(x_2) = 0{,}89$; $S_e(x_3) = 0{,}39$; $S_e(y) = 0{,}89$
2. $S_{y.x_1} = 3{,}2$; $S_{y.x_2} = 3{,}67$; $S_{y.x_3} = 3{,}81$
3. $y' = 0{,}8x_1$; $y' = 0{,}8x_2$; $y' = 1{,}2x_3$
4. $Z_{y'} = 0{,}60Z_{x_1} + 0{,}40Z_{x_2} + 0{,}30Z_{x_3}$
5. 61%
6.

	y	x_1	x_2	x_3
y	16			
x_1	7,2	9		
x_2	3,2	0,0	4	
x_3	1,2	0,0	0,0	1

22.1.1. 15
1.2. 0,80; 0,875; 0,714; 0,60
2.1. 13
2.2. 0,73; 1; 0,43; 0,43
3.1. 3; 2

4 Análisis de los ítems

Hasta ahora nos hemos ocupado de las propiedades del test considerado globalmente, de su capacidad para discriminar entre las personas, de su fiabilidad y de las evidencias de validez. Ahora bien, en el proceso real de construcción de un test se empieza por elaborar un número elevado de ítems, dos o tres veces más de los que el test tendrá finalmente, aplicar esos ítems a una muestra de personas semejantes a aquellas a las que el test irá destinado y descartar los que no sean pertinentes. La cuestión es cómo saber qué ítems son pertinentes, objetivo central del análisis de ítems.

Se entiende por análisis de ítems el estudio de aquellas propiedades de los ítems que están directamente relacionadas con las propiedades del test y, en consecuencia, influyen en ellas. En palabras de Lord y Novick (1968), el requerimiento básico de un parámetro de un ítem es que tenga una relación clara con algún parámetro interesante del test total. Se previene, por tanto, al lector contra retahílas de descriptores de los ítems que a veces aparecen en los textos sin hacer referencia alguna a su incidencia en los parámetros del test. Son perfectamente inútiles, pues de ellos no se colige ninguna inferencia directa sobre el test.

Aquí se tratarán los tres índices más relevantes:

— índice de dificultad,
— índice de discriminación,
— índice de validez,

y se especificarán sus relaciones con los parámetros del test considerado globalmente. Para un tratamiento más completo, puede verse Muñiz et al. (2005a).

Además, se incluyen en este apartado otras consideraciones complementarias para el estudio de los ítems, tales como el análisis de las alternativas incorrectas, la corrección del azar, la calificación del conocimiento parcial y algunas técnicas para la evaluación del funcionamiento diferencial.

1. ÍNDICE DE DIFICULTAD

Se entiende por índice de dificultad (ID) de un ítem la proporción de personas que lo aciertan de aquellas que han intentado resolverlo:

$$ID = \frac{A}{N} \qquad [4.1]$$

donde

A: Número de personas que aciertan el ítem.
N: Número de personas que han intentado resolver el ítem.

El valor del índice de dificultad está directamente relacionado con la media del test:

$$\bar{X} = \sum_{i=1}^{n} ID_i \qquad [4.2]$$

En palabras, la media del test es igual a la suma de los índices de dificultad de los ítems.

Los cálculos de la tabla 4.1 permiten ilustrar la citada igualdad.

TABLA 4.1

Personas	Ítems				Puntuación total
	1	2	3	4	
A	0	1	1	1	3
B	1	0	1	0	2
C	1	1	0	0	2
D	1	1	1	1	4
E	0	1	0	0	1
ID_i	3/5	4/5	3/5	2/5	12

Al índice de dificultad sería semánticamente más apropiado denominarlo «índice de facilidad», pues, a medida que aumenta, indica que el ítem es más fácil, no más difícil. En la tabla anterior, por ejemplo, el ítem más fácil es el segundo, que es acertado por cuatro de las cinco personas; sin embargo, su índice de dificultad es el mayor (4/5).

Nótese también que en muchos test no tiene ningún sentido hallar el índice de dificultad de los ítems; por ejemplo, en test dirigidos a evaluar aspectos de personalidad, en los que los ítems no son fáciles ni difíciles.

Una seria limitación de este índice de dificultad de la teoría clásica es su dependencia directa de la muestra de personas en la que se calcula, es decir, el índice de dificultad no constituye una propiedad intrínseca del ítem, su valor depende del tipo de personas a las que se aplique. Si son muy competentes, resultará un ítem fácil, lo acertarán muchos. Si, por el contrario, son incompetentes, el mismo ítem resultará difícil. A nivel práctico, la teoría clásica mitiga este inconveniente calculando el índice de dificultad en muestras similares en competencia con aquellas en las que se van a usar posteriormente los ítems. Ahora bien, este recurso resulta poco convincente a nivel teórico para una teoría de la medición psicológica medianamente rigurosa, donde sería de esperar que las propiedades de los instrumentos de medida no dependiesen de los objetos medidos. Una solución adecuada a este problema la proporcionarán los modelos de teoría de respuesta a los ítems, como se verá más adelante.

Cuando los ítems son de elección múltiple y, en consecuencia, es posible acertarlos por mero azar, el índice de dificultad conviene calcularlo corrigiendo los efectos del azar mediante la fórmula clásica que se presenta a continuación, aunque otras son también posibles:

$$ID = \frac{A - \frac{E}{K-1}}{N} \qquad [4.3]$$

donde

A: Número de personas que aciertan el ítem.
E: Número de personas que fallan el ítem.
K: Número de alternativas del ítem.
N: Número de personas que intentan resolver el ítem.

La varianza de un ítem puede expresarse en términos de su índice de dificultad, puesto que para una variable dicotómica *j*: $\sigma_j^2 = P_j Q_j$, donde P_j sería aquí la proporción de personas que aciertan el ítem, es decir, el índice de dificultad, y $Q_j = (1 - P_j)$. La varianza será máxima para los valores medios de P_j; en otras palabras, la dificultad media de los ítems maximiza su varianza.

2. ÍNDICE DE DISCRIMINACIÓN

Se dice que un ítem tiene poder discriminativo si distingue, discrimina, entre aquellas personas que puntúan alto en el test y las que puntúan bajo, es decir, si discrimina entre los eficaces en el test y los ineficaces. En consecuencia, el índice de discriminación se define como la correlación entre las puntuaciones de las personas en el ítem y sus puntuaciones en el test.

Cuál haya de ser el tipo de correlación a utilizar dependerá de las características de las variables a correlacionar, en nuestro caso el ítem y el test. Aquí se ilustrarán cuatro de las correlaciones más habituales dados los formatos que suelen adoptar más frecuentemente los ítems y los tests, pero otras muchas son posibles, y en cada caso habrá que elegir la más adecuada. Una interesante discusión acerca de la elección de correlación puede verse en Carroll (1961).

2.1. Cálculo

Veamos a continuación cuatro posibles coeficientes de correlación para la estimación del índice de discriminación:

— correlación biserial-puntual,
— correlación biserial,
— coeficiente *phi,*
— correlación tetracórica,

para luego analizar las conexiones entre el índice de discriminación y los parámetros del test.

Correlación biserial-puntual (ρ_{bp})

La correlación biserial-puntual es una mera aplicación de la correlación de Pearson cuando una de las variables es dicotómica y la otra cuantitativa continua, o eventualmente discreta (Amón, 1984). Suele usarse con frecuencia para calcular el índice de discriminación, dado que es habitual que los ítems sean dicotómicos (o se aciertan o se fallan), y el test constituya una medida cuantitativa discreta. La fórmula de la correlación de Pearson en estas circunstancias viene dada por:

$$\rho_{bp} = \frac{\mu_p - \mu_x}{\sigma_x}\sqrt{\frac{p}{q}} \qquad [4.4]$$

donde

μ_p: Media en el test de las personas que aciertan el ítem.
μ_x: Media del test.
σ_x: Desviación típica del test.
p: Proporción de personas que aciertan el ítem.
q: $(1 - p)$.

Ejemplo

Calcular el índice de discriminación del tercer ítem de la tabla 4.1 del apartado precedente.

En primer lugar, para realizar los cálculos indicados por la fórmula [4.4], a la puntuación total del test (X) hay que descontarle el ítem cuyo índice de discriminación se pretende hallar ($X - j$); de lo contrario, una de las variables a correlacionar (el ítem) estaría impropiamente incluida en la otra (el test). Véase la nota que sigue al ejemplo.

TABLA 4.2

Sujetos	Ítems				Total	
	1	2	3	4	X	$(X - j)$
A	0	1	1	1	3	2
B	1	0	1	0	2	1
C	1	1	0	0	2	2
D	1	1	1	1	4	3
E	0	1	0	0	1	1

$$\bar{X}_p = \frac{2+1+3}{3} = 2$$

$$\bar{X}_x = \frac{2+1+2+3+1}{5} = 1{,}8$$

$$S_x^2 = \frac{2^2+1^2+2^2+3^2+1^2}{5} - (1{,}8)^2 = 0{,}56$$

$$S_x = \sqrt{0{,}56} = 0{,}748$$

$$p = \frac{3}{5} = 0{,}60$$

$$q = 1 - 0{,}60 = 0{,}40$$

$$r_{bp} = \frac{2 - 1{,}8}{0{,}748}\sqrt{\frac{0{,}60}{0{,}40}} = 0{,}32$$

Nota. Si al calcular la correlación ítem-test no se descontase de este, como se hizo en el ejemplo, las puntuaciones correspondientes al ítem, se estaría elevando impropia y espuriamente la correlación, pues estrictamente no se estaría correlacionando el ítem con el resto de los ítems (test), sino con un test que incluiría también el ítem en cuestión. En suma, se estaría correlacionando una variable (test) con parte de ella (ítem). Bien es verdad que cuando el test consta de un número elevado de ítems, este efecto puede ser de poca relevancia empírica, pero ello no legitima, claro está, su incorrección.

Lo más sencillo es calcular la correlación, como se ha hecho en el ejemplo, descontando el ítem. No

obstante, si por cualquier razón se tiene la correlación ítem-test sin descontar los efectos del ítem, puede utilizarse la siguiente fórmula de corrección para obtener la correlación pertinente:

$$\rho_{j(x-j)} = \frac{\rho_{jx}\sigma_x - \sigma_j}{\sqrt{\sigma_j^2 + \sigma_x^2 - 2\rho_{jx}\sigma_j\sigma_x}} \qquad [4.5]$$

donde

$\rho_{j(x-j)}$: Correlación entre el ítem j y el test tras descontar el ítem $(x - j)$.
ρ_{jx}: Correlación ítem-test cuando el ítem está incluido en el test.
σ_x: Desviación típica del test.
σ_j: Desviación típica del ítem.

La obtención de [4.5] es inmediata a partir de la fórmula general de la correlación de Pearson:

$$\rho_{j(x-j)} = \frac{Ej(x-j)}{\sigma_j\sigma_{(x-j)}} = \frac{Ejx - Ej^2}{\sigma_j\sqrt{\sigma_x^2 + \sigma_j^2 - 2\rho_{jx}\sigma_j\sigma_x}}$$

Teniendo en cuenta el valor de la esperanza matemática y simplificando:

$$\rho_{j(x-j)} = \frac{\rho_{jx}\sigma_x - \sigma_j}{\sqrt{\sigma_x^2 + \sigma_j^2 - 2\rho_{jx}\sigma_j\sigma_x}}$$

Apliquemos lo dicho a nuestro ejemplo. En primer lugar vamos a calcular la correlación ítem-test sin excluir el ítem del test, posteriormente aplicaremos al valor obtenido la corrección propuesta y el resultado deberá ser el mismo que el obtenido al principio cuando habíamos descontado el ítem.

$$\bar{X}_p = \frac{3+2+4}{3} = 3$$

$$\bar{X}_x = \frac{3+2+2+4+1}{5} = 2{,}4$$

$$S_x^2 = \frac{3^2+2^2+2^2+4^2+1^2}{5} - (2{,}4)^2 = 1{,}04$$

$$S_x = \sqrt{1{,}04} = 1{,}02$$

$$p = \frac{3}{5} = 0{,}6$$

$$q = 1 - 0{,}6 = 0{,}4$$

$$r_{bp} = \frac{3 - 2{,}4}{1{,}02}\sqrt{\frac{0{,}6}{0{,}4}} = 0{,}72$$

Aplicando la corrección propuesta en [4.5]:

$$r_{j(x-j)} = \frac{(0{,}72)(1{,}02) - 0{,}49}{\sqrt{0{,}24 + 1{,}04 - 2(0{,}72)(0{,}49)(1{,}02)}} = 0{,}32$$

que, efectivamente, es el mismo resultado que el obtenido en principio.

Correlación biserial (ρ_b)

Si una de las variables a correlacionar, que en las presentes circunstancias suele ser el ítem, no es dicotómica por naturaleza, pero por alguna razón se dicotomiza y se asume que bajo esa dicotomización subyace una variable continua distribuida según la curva normal, puede usarse la correlación biserial (ρ_b) para estimar el índice de discriminación. La situación citada se da con cierta frecuencia, por ejemplo, cuando se dicotomizan ítems a pesar de admitir una gradación de respuestas. Si se puede evitar, es desaconsejable la dicotomización, puesto que con ella siempre se pierde información, reduciendo la escala de medición a solo dos categorías.

$$\rho_b = \frac{\mu_p - \mu_x}{\sigma_x}\frac{p}{y} \qquad [4.6]$$

donde, asumiendo que la variable dicotomizada es el ítem:

μ_p: Media en el test de las personas que aciertan el ítem.
μ_x: Media del test.
σ_x: Desviación típica del test.
p: Proporción de personas que aciertan el ítem.
y: Ordenada correspondiente al valor de la puntuación típica en la curva normal que

deja por debajo un área igual a p. (Los valores de y pueden obtenerse en la tabla II.)

Lo dicho para la correlación biserial-puntual sigue siendo válido aquí en líneas generales, si bien hay que tener en cuenta que, a diferencia de ρ_{bp}, la correlación biserial no es una mera aplicación de la correlación de Pearson, sino una estimación de ella. De modo que podrían obtenerse valores mayores que 1, especialmente si alguna de las variables es platicúrtica o bimodal. Si se sospecha una distribución normal dudosa, es más seguro utilizar ρ_{bp}, máxime si las correlaciones se van a usar en análisis de regresión o factoriales.

La relación entre ρ_{bp} y ρ_b viene dada por:

$$\rho_{bp} = \rho_b \sqrt{\frac{pq}{y}}$$

donde p, q e y tienen la significación ya citada.

Coeficiente *phi* (ϕ)

Si las variables a correlacionar, en nuestro ítem y test, son ambas dicotómicas, un coeficiente adecuado para estimar el índice de discriminación viene dado por el coeficiente ϕ, que es una mera aplicación del coeficiente de correlación de Pearson.

Correlación tetracórica

Si ambas variables (ítem y test) están dicotomizadas y se asumen distribuidas normalmente, la correlación tetracórica es el coeficiente adecuado para estimar el índice de discriminación.

Para un análisis comparativo de los coeficientes citados y de su comportamiento en situaciones concretas de selección de ítems, véase Lord y Novick (1968). Una exposición detallada y clara de su cálculo, propiedades y significación estadística puede consultarse en Amón (1984), San Martín y Pardo (1989).

Índice basado en las proporciones de aciertos

Cabe añadir, finalmente, un método elemental y directo de acercarse al índice de discriminación no basado en la correlación ítem-test. Este índice (d) es la diferencia entre la proporción de personas competentes que aciertan el ítem (P_c) y la proporción de incompetentes que también lo aciertan, entendiendo por competentes aquellos que puntúan en el test por encima de la mediana, e incompetentes por debajo. Pueden utilizarse grupos más extremos, siendo clásicos el 27% superior e inferior sugeridos por Kelley (1939).

$$d = P_c - P_i \qquad [4.7]$$

donde

P_c: Proporción de personas competentes en el test que aciertan el ítem.
P_i: Proporción de personas incompetentes en el test que también aciertan el ítem.

La interpretación de d es obvia: la capacidad discriminativa del ítem aumenta a medida que d se aleja de cero, bien sea hacia 1 o hacia −1. En el caso extremo de que fuese 1, significaría que todos los competentes aciertan el ítem ($P_c = 1$) y todos los incompetentes lo fallan ($P_i = 0$); la discriminación es perfecta. En el caso de −1, $P_c = 0$ y $P_i = 1$, sería el caso paradójico en el que todos los incompetentes lo aciertan y todos los competentes lo fallan; luego el ítem también discrimina perfectamente, pero habría que tener cuidado a la hora de la interpretación.

Cuando el acceso a los ordenadores era limitado, se utilizaban un sinfín de tablas para hacer estimaciones de las correlaciones a partir de las proporciones de aciertos. Hoy esto carece de sentido, ya que todos los cálculos están implementados en numerosos programas, tanto comerciales como de libre acceso.

2.2. Relación con algunos parámetros del test

a) *Variabilidad*

Una medida de la capacidad discriminativa de un test es la variabilidad de las puntuaciones obtenidas en él por las personas, es decir, su desviación típica (σ_X). Cuando esta es cero ($\sigma_X = 0$), no hay

discriminación alguna, todas las personas sacan la misma puntuación; el test no distingue, no discrimina entre unas personas y otras. Ahora bien, la desviación típica del test está íntimamente relacionada con el índice de discriminación de los ítems:

$$\sigma_x = \sum_{j=1}^{n} \sigma_j \rho_{jx} \qquad [4.8]$$

donde

σ_X: Desviación típica del test.
σ_j: Desviación típica del ítem *j*.
ρ_{jX}: Índice de discriminación del ítem *j*.

Efectivamente, sean X las puntuaciones en el test y x_j las obtenidas en los n ítems, con $X = \sum x_j$, la varianza de X, o, lo que es lo mismo, la covarianza consigo misma, vendrá dada por:

$$\begin{aligned}\sigma_X^2 &= \sigma(X, X) = \sigma(X, \sum x_1) = \\ &= \sigma(X, x_1 + x_2 + \cdots + x_n) = \\ &= \sigma(X, x_1) + \sigma(X, x_2) + \sigma(X, x_3) + \cdots + \sigma(X, x_n) = \\ &= \sum \sigma(X, x_j)\end{aligned}$$

sustituyendo la covarianza por su valor:

$$\sigma_X^2 = \sum \sigma_X \sigma_j \rho_{jX}$$

y simplificando:

$$\sigma_X = \sigma_j \rho_{jX}$$

que es la fórmula propuesta en [4.8].

Por tanto, la capacidad discriminativa de un test depende directamente de la desviación típica de sus ítems (σ_j) y de la correlación de estos con el test total (ρ_{jX}), es decir, de su índice de discriminación. (En la literatura psicométrica suele denominarse índice de fiabilidad del ítem *j* al producto $\sigma_j \rho_{jX}$.)

Si los ítems son dicotómicos, su desviación típica viene dada por:

$$\sigma_j = \sqrt{P_j Q_j} = \sqrt{P_j (1 - P_j)}$$

donde P_j es la proporción de sujetos que aciertan el ítem *j*, esto es, su índice de dificultad. Sustituyendo en [4.8]:

$$\sigma_X = \sum_{j=1}^{n} \sqrt{P_j (1 - P_j)} \rho_{jX} \qquad [4.9]$$

En esta fórmula queda claro que para maximizar la capacidad discriminativa de un test hay que tener en cuenta conjuntamente el índice de dificultad de los ítems (P_j) y el índice de discriminación (ρ_{jX}).

Adviértase que un error muy frecuente es considerar estos dos parámetros por separado y afirmar sin más que un ítem contribuye a la discriminación global del test cuando él mismo tiene un gran poder discriminativo, esto es, una varianza máxima, lo cual ocurre para $P_j = 0{,}50$. Ello no es estrictamente cierto, pues si $\rho_{jX} = 0$, aun con $P_j = 0{,}50$, ese ítem no contribuye en absoluto a la capacidad discriminativa del test, puesto que:

$$\sigma_X = \sqrt{P_j (1 - P_j)} \rho_{jX} = \sqrt{0{,}50(1 - 0{,}50)}(0) = 0$$

En suma, que un ítem sea muy discriminativo no implica automáticamente que contribuya a la misma discriminación hecha por el test; hay que considerar además su correlación con el test, es decir, su índice de discriminación.

b) *Fiabilidad*

La fiabilidad del test también puede expresarse en función de la varianza de los ítems (σ_j^2) y de sus índices de discriminación (ρ_{jX}), para lo cual se sustituye el valor de σ_X dado por [4.9] en la fórmula del coeficiente α:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_j^2}{[\sum \sigma_j \rho_{jx}]^2}\right) \qquad [4.10]$$

O en el caso de que los ítems sean dicotómicos:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum P_j (1 - P_j)}{[\sum \rho_{jx} \sqrt{P_j (1 - P_j)}]^2}\right) \qquad [4.11]$$

En suma, los parámetros de los tests, poder discriminativo (σ_X) y fiabilidad (α), pueden expresarse en términos del índice de dificultad de los ítems (P_j) y de su índice de discriminación (ρ_{jX}).

3. ÍNDICE DE VALIDEZ

Se denomina «índice de validez» de un ítem a su correlación con el criterio. Sobre qué correlación utilizar solo cabe repetir lo dicho en el apartado anterior para el caso del índice de discriminación: dependerá de la naturaleza de las variables a correlacionar, que aquí son ítem y criterio. Como ocurría con el índice de discriminación, también para la validez las correlaciones más frecuentes son la biserial puntual, biserial, *phi* y tetracórica. Su cálculo es idéntico, si bien ahora no existe el problema adicional de que el ítem esté en ocasiones incluido en el criterio, como podía ocurrir al correlacionar ítem-test, según se ha visto.

3.1. Relación con los parámetros del test

Sea cual sea la correlación utilizada, el valor del índice de validez de un ítem indicará en qué grado el ítem está conectado con la variable que el test intenta predecir (criterio). Desde este punto de vista, la validez global de un test se verá incrementada en la medida en que sus ítems tienen índices de validez elevados. En concreto (véase Gulliksen, 1950), la conexión entre el índice de validez de los ítems y el coeficiente de validez del test viene dada por:

$$\rho_{xy} = \frac{\sum_{j=1}^{n} \sigma_j \rho_{jY}}{\sum_{j=1}^{n} \sigma_j \rho_{jX}} \qquad [4.12]$$

donde

ρ_{xy}: Coeficiente de validez del test.
n: Número de ítems del test.
σ_j: Desviación típica del ítem *j*.
ρ_{jY}: Índice de validez del ítem *j*.
ρ_{jX}: Índice de discriminación del ítem *j*.

Si los ítems fuesen dicotómicos, entonces:

$$\sigma_j = \sqrt{P_j Q_j} = \sqrt{P_j(1 - P_j)}$$

y sustituyendo, la fórmula anterior [4.12] puede expresarse del siguiente modo:

$$\rho_{xy} = \frac{\sum_{j=1}^{n} \rho_{jY} \sqrt{P_j(1 - P_j)}}{\sum_{j=1}^{n} \rho_{jX} \sqrt{P_j(1 - P_j)}} \qquad [4.13]$$

donde todos los símbolos son los mismos que en [4.12] y P_j es el índice de dificultad del ítem *j*.

La fórmula anterior [4.13] es de suma importancia, pues expresa el coeficiente de validez del test en función de los tres parámetros de los ítems: dificultad (P_j), discriminación (ρ_{jX}) y validez (ρ_{jY}).

Por otra parte, recuérdese que según [4.11] la fiabilidad del test (a) podía expresarse en función de la dificultad y discriminación de los ítems:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum P_j(1 - P_j)}{[\sum \rho_{jX} \sqrt{P_j(1 - P_j)}]^2}\right) \qquad [4.11]$$

Una paradoja clásica

Al contemplar ambas expresiones, [4.11] y [4.13], aparece una conocida paradoja de la teoría clásica de los tests, a saber, según [4.11], si se desea maximizar la fiabilidad del test, habrá que elegir ítems con índices de discriminación elevados (ρ_{jX}), supuestas iguales otras cosas, pero, según [4.13], al elegir índices de discriminación elevados se rebaja el coeficiente de validez del test, ya que según [4.13] la validez aumenta cuanto mayores sean los índices de validez de los ítems y menores sean sus índices de discriminación.

Por tanto, si se desea maximizar la validez del test, han de elegirse ítems con elevados índices de validez, pero en contrapartida el test resultante tal vez no tenga en ese caso una elevada fiabilidad. Nótese que en este contexto de fiabilidad se refiere a la consistencia interna (α), cuyas relaciones con otros tipos de fiabilidad ya se trataron en su momento. El

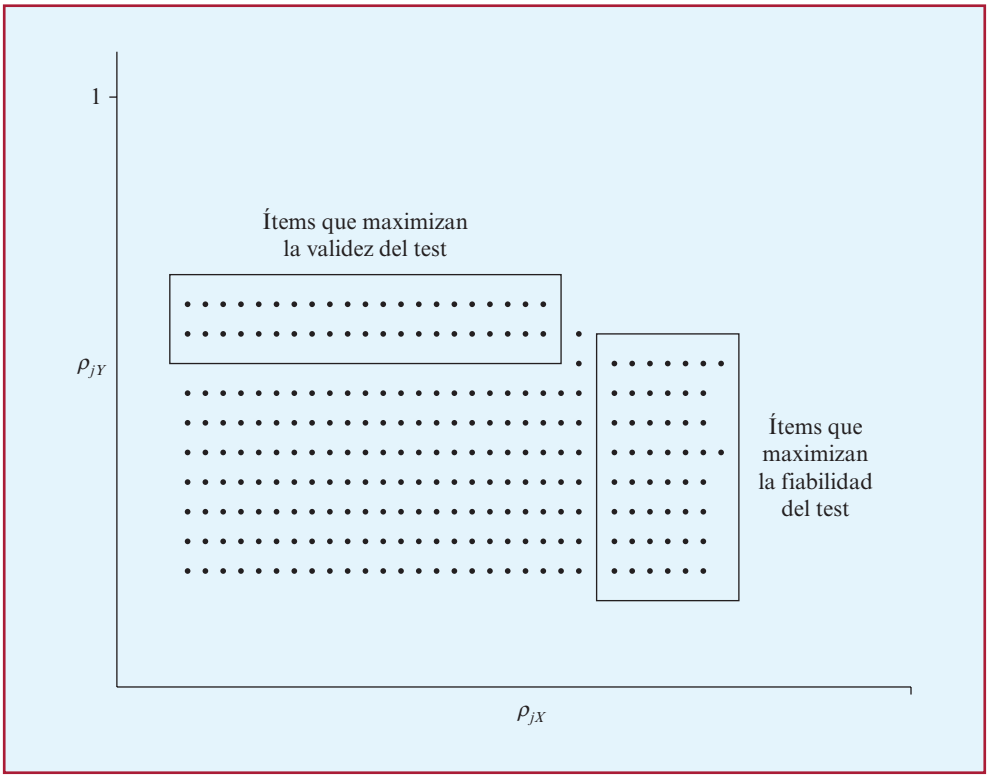


Figura 4.1.—Paradoja clásica.

criterio a maximizar está en función de los intereses del investigador y de las características de los objetivos a medir. Véase a modo de ilustración la figura 4.1 en la que en abscisas se representa el valor del índice de discriminación, y en ordenadas, el del índice de validez.

Comentarios finales

No debe confundirse el índice de validez con lo que a veces se denomina «evidencias de validez factorial de los ítems», refiriéndose a la estructura factorial de los ítems tras someterlos a un análisis factorial, y que da una idea de la estructura interna del test, esto es, de si todos los ítems configuran uno o más factores. Recuérdese que otra medida de la cohesión interna de los ítems es el coeficiente *alfa,* aunque un *alfa* elevado no garantiza una estructura unifactorial, como a veces se afirma.

Otro concepto de interés relativo a los ítems es el de su posible ponderación. En ocasiones, puede tener interés y ser aconsejable en función de los objetivos perseguidos que no todos los ítems tengan la misma ponderación o peso a la hora de contribuir a la puntuación total del test. Caso típico y seguramente familiar al lector sería el de un examen o prueba objetiva en la que no todas las preguntas (ítems) valen lo mismo, debido a cierto criterio de ponderación establecido por el profesor.

Algunas técnicas estadísticas como la regresión múltiple pueden ser de gran utilidad a la hora de establecer las citadas ponderaciones.

Por último, señalar que en el proceso de seleccionar los ítems que van a constituir el test definitivo hay que tener presentes dos cuestiones. En primer lugar, que al elegir aquellos ítems con índices de discriminación y de validez elevados se está capitalizando el error, es decir, si dichos índices se calculasen en una segunda muestra de personas, seguramente tenderían a bajar. En segundo lugar, el índice de discriminación de un ítem es la correlación ítem-test. Depende, por tanto, no solo del ítem sino del resto de los ítems que

constituyan el test, luego al descartar algunos de ellos tenderá a variar, dado que con el descarte varía la variable «test». Esta incidencia será menor cuantos menos ítems se descarten, siendo aconsejable hacer la selección de ítems en varios pasos o etapas, no de una vez, descartando un bloque de ítems de cada vez y recalculando los parámetros tras el descarte.

4. ANÁLISIS DE LAS ALTERNATIVAS INCORRECTAS

Además del cálculo de los parámetros de los ítems citados, conviene indagar la distribución de las respuestas de las personas a las alternativas incorrectas de los ítems, cuando estos son de elección múltiple, por si ello pudiera explicar la inadecuabilidad de alguno de ellos. Los programas de ordenador para analizar los ítems desde el punto de vista de la teoría clásica facilitan esta información acerca de la proporción de personas que contestan cada alternativa. Por ejemplo, un índice de discriminación bajo puede venir a veces explicado porque una de las alternativas falsas «atrae» por igual y masivamente a los competentes e incompetentes en el test, y tal vez el mero hecho de cambiarla por otra más adecuada podría ser suficiente para incrementar el índice. En otros casos se observa que ciertas alternativas no son elegidas por nadie, ni competentes ni incompetentes, por lo que no contribuyen en nada a la discriminación, etc.

Véase el ejemplo numérico que sigue (tabla 4.3), en el que aparecen las respuestas de una muestra de 200 personas a un ítem con cinco alternativas, siendo la tercera (*C*) la alternativa correcta. Se denomina competentes al 50% que están por encima de la mediana del test, e incompetentes a los que quedan por debajo.

TABLA 4.3

Ítem	Alternativas					
	A	B	C	D	E	
Competentes	5	15	70	10	0	100
Incompetentes	15	65	20	0	0	100
	20	80	90	10	0	200

Las alternativas *D* y *E* parece claro que habría que sustituirlas por otras; la *D* por alguna razón «atrae» más a los competentes que a los incompetentes, mientras que la *E* no recibe ninguna respuesta, seguramente por demasiado obvia. No es nada infrecuente encontrarse con alternativas como la *D* que, paradójicamente, son más elegidas por los que puntúan alto en el test que por los incompetentes. Aunque la explicación ha de buscarse en cada caso particular, suele ocurrir con ese tipo de alternativas que contienen información de un nivel elevado que problematiza a los que más saben, pasando desapercibida, sin embargo, para los menos competentes.

4.1. Número óptimo de alternativas

Una de las primeras preguntas que asaltan al que intenta construir un instrumento de medida con ítems de elección múltiple se refiere a cuál es el número óptimo de alternativas que deben tener los ítems. La respuesta parece simple: cuantas más mejor, pues al aumentar su número se reducirá la probabilidad de aciertos al azar. Ahora bien, la pregunta se puede sofisticar un poco más. Supóngase un test de 10 ítems con cinco alternativas cada uno. Para intentar responderlo, las personas tendrán que leer 50 frases, que le ocuparán un tiempo determinado, pongamos 50 minutos. Otro test que mide la misma variable está compuesto de 25 ítems, con dos alternativas por ítem; luego el tiempo exigido es el mismo, 50 minutos. ¿Cuál de los dos tests será preferible? O, en otras palabras, asumido un cierto tiempo límite, ¿cuál es el número óptimo de alternativas por ítem? Dadas las claras implicaciones prácticas para la construcción de test, esta pregunta ha sido abordada clásicamente por numerosos autores (Ebel, 1951; Grier, 1975, 1976; Lord, 1977, 1980; Tversky, 1964).

Trabajos empíricos pioneros citados por Lord (1980) parecen indicar que ítems con dos o tres alternativas dan fiabilidades tan buenas o mejores que los ítems con cuatro o cinco alternativas. Según Tversky (1964), en un elegante razonamiento matemático, el número óptimo de alternativas sería tres (exactamente 2,72, base de los logaritmos neperianos). Las conclusiones de Grier (1975, 1976) también están a favor de tres alternativas, seguidas en

pertinencia por dos alternativas. Lord (1977, 1980), tras una buena revisión y comentario de las aproximaciones precedentes, deriva una fórmula según la cual el número óptimo de alternativas vendría dado por:

$$A = 1 + \frac{1}{\sqrt{(1-r)p}} \qquad [4.14]$$

donde

p: Índice de dificultad del ítem.
r: Correlación entre dos ítems equivalentes con infinitas alternativas. Sería la fiabilidad teórica del ítem cuando no existe la posibilidad de acierto al azar, al tener infinitas alternativas.

Por ejemplo, para un ítem de dificultad media ($p = 0{,}50$) y $r = 0{,}50$, el número óptimo de alternativas según [4.14] vendría dado por:

$$A = 1 + \frac{1}{\sqrt{(1-0{,}50)(0{,}50)}} = 3$$

Sin embargo, para otros valores de p y r, A puede tomar valores bien distintos. Compruébelo el lector.

La fórmula de Lord (1980), más que contradecir los resultados de Grier y de Tversky, los matiza, si bien los supuestos en los que se basa es poco probable que se cumplan estrictamente en la práctica (véase Lord, 1980, pp. 108-109). A la vista de los datos, el número de tres alternativas aparece como el más recomendable, y así lo confirma el metaanálisis de Rodríguez (2005). No obstante, los resultados han de tomarse con cierta precaución, pues los datos de Budescu y Nevo (1985) no confirman la hipótesis de proporcionalidad en la que se basan estas estimaciones, hipótesis que se refiere a la asunción de que el tiempo total para contestar el test es proporcional al número de ítems y de alternativas por ítem. Para una buena revisión de la problemática relativa al número de alternativas, véase Delgado y Prieto (1998).

Otra cuestión de alto interés es si la modificación del número de alternativas tiene el mismo efecto sobre la eficacia del test para los distintos niveles de competencia de las personas, pues, como es bien sabido, las personas incompetentes, en especial las muy incompetentes, lo suelen hacer peor que si contestasen al azar, debido, al parecer, a que son «seducidos» por alternativas falsas plausibles para ellos, cosa que no les ocurriría de seguir los azarosos dictados de una moneda sin lastrar. Los modelos de teoría de respuesta a los ítems permiten analizar de una manera adecuada esta incidencia para los distintos niveles.

5. CORRECCIÓN DEL AZAR

Normalmente la puntuación de una persona en un test se obtiene mediante una combinación lineal de sus puntuaciones en los ítems, bien sea sumando estas o sumándolas tras ponderarlas de algún modo. En la gran mayoría de los casos la puntuación total es la suma de las obtenidas en los ítems. Un caso especial surge cuando los ítems son de elección múltiple, en cuyo caso, aun sin conocer la respuesta correcta, las personas pueden acertar por mero azar. En estos casos la fórmula más clásica para la corrección de los aciertos azar viene dada por:

$$P = A - \frac{E}{n-1} \qquad [4.15]$$

donde

A: Número de aciertos.
E: Número de errores.
n: Número de alternativas del ítem.

Ejemplo

Los 100 ítems que componen un test tienen cuatro alternativas cada uno, de las cuales solo una es correcta. Si un sujeto contestó correctamente 60 ítems, falló 27 y no respondió a 13, ¿qué puntuación se le asignará una vez corregidos los efectos del azar?

$$P = 60 - \frac{27}{4-1} = 51$$

La fórmula [4.15] se basa en ciertos supuestos que de no cumplirse la invalidan. Asume que las respuestas correctas provienen o bien de que la persona las conoce y, por tanto, responde adecuadamente, o bien de que, aun sin conocerlas, responde al azar y acierta. En el caso de los errores se presume que las personas desconocían el ítem, respondieron al azar y fallaron. Asimismo, se asume que cuando no se conoce la respuesta a un ítem y, por tanto, se contesta al azar, todas las alternativas del ítem son equiprobables. Nótese que esta última asunción es especialmente débil, pues generalmente de las alternativas falsas de un ítem suele conocerse alguna, descartándola, con lo que se viola el citado supuesto, al aumentar la probabilidad de acierto de la alternativa correcta.

Cuando se utilice [4.15] es imprescindible hacérselo saber a las personas evaluadas para unificar en lo posible su conducta a la hora de correr riesgos en caso de duda, si bien dicha unificación nunca será perfecta, puesto que depende, entre otras cosas, de las características orécticas de las personas.

Nótese que si, como en ocasiones se hace, se instruye a las personas para que no omitan ningún ítem (práctica, por otra parte, poco recomendable), las puntuaciones obtenidas correlacionan perfectamente ($r_{xy} = 1$) con las corregidas según la fórmula [4.15], por lo que su uso tiene poco sentido, o más bien ninguno. En la literatura psicométrica existe un amplio y especializado capítulo de trabajos acerca de la supuesta conducta de las personas ante los ítems, posibles técnicas de puntuación, casuística, influencia en los distintos parámetros del test, etc. El lector interesado puede consultarlo en Albanese (1986), Angoff y Schrader (1984, 1986), De Finetti (1965), Diamond y Evans (1973), Frary (1980), García-Pérez (1987, 1989), Gibbons et al. (1979), Hutchinson (1982), Lord (1975, 1980), Lord y Novick (1968), Rowley y Traub (1977), Schmittlein (1984), Traub y Hambleton (1972), Wainer (1983), Wilcox (1983, 1985), entre otros muchos.

Obtención de la fórmula de corrección

Según los supuestos citados, si cada ítem tiene n alternativas con solo una correcta, la probabilidad de acertarlo al azar será:

$$P(A) = \frac{1}{n}$$

y la probabilidad de error:

$$P(E) = \frac{(n-1)}{n}$$

Ahora bien, cuando se corrige el test, solo se tienen dos datos del sujeto: el número de aciertos (A) y el número de errores (E). Lo que intentamos hacer corrigiendo el azar es restar de los aciertos (A) aquellos que se deben al azar (A_a), esto es, calcular:

$$P = A - A_a$$

Sería muy fácil si supiésemos las respuestas que el sujeto dio al azar (R_a), pues los aciertos al azar (A_a) serían igual al número de respuestas al azar (R_a) dividido entre n:

$$A_a = \frac{R_a}{n}$$

Ahora bien, sí sabemos que:

$$E = R_a \frac{(n-1)}{n}$$

despejando R_a:

$$R_a = \left[\frac{n}{(n-1)}\right] E$$

sustituyendo:

$$A_a = \frac{\left[\dfrac{n}{(n-1)}\right] E}{n} = \frac{nE}{(n-1)n} = \frac{E}{(n-1)}$$

de donde:

$$P = A - A_a = A - \frac{E}{(n-1)}$$

que es la fórmula propuesta en [4.15].

Prohibición de las omisiones

Como se ha señalado, si las personas son instruidas para responder a todos los ítems, obligándoles a contestar al azar cuando no conocen la respuesta, entonces:

$$E = N - A$$

siendo N el número de ítems. Sustituyendo en [4.15]:

$$P = A - \frac{N - A}{n - 1} = \frac{(n-1)A - (N - A)}{n - 1} =$$

$$= \frac{nA - A - N + A}{n - 1} = \frac{n}{n - 1} A - \frac{N}{n - 1}$$

Ahora bien, P es una función lineal de A, puesto que $n/(n - 1)$ y $N/(n - 1)$ son constantes: luego su correlación es 1 ($r_{PA} = 1$).

En suma, en esta situación, como se apuntó al principio, daría igual escalar a las personas en función de la puntuación corregida P que en función del número de aciertos A.

Nótese también que si bajo las instrucciones de no omitir ningún ítem alguna persona sí lo hace, su puntuación global ha de ser corregida para compensar las respuestas correctas que habría obtenido al azar de haber rellenado todos los ítems. La corrección, siguiendo la lógica anterior, consistiría en sumar a sus aciertos el número de omisiones (O) dividido entre n:

$$P = A + \frac{O}{n}$$

6. CALIFICACIÓN DEL CONOCIMIENTO PARCIAL

Dos personas que aciertan determinado ítem tal vez no posean el mismo grado de conocimiento acerca de él, y lo mismo puede decirse si lo fallan. La psicometría ha tratado de antiguo de averiguar de alguna manera gradaciones en la calificación de las respuestas, es decir, se ha tratado de calificar por diversos caminos el conocimiento parcial o incompleto que las personas poseen de los ítems, y no quedarse escuetamente en el acierto/fallo. Se han propuesto acercamientos al problema muy diversos, entre los que cabe destacar:

— Juicios de seguridad.
— Responder-hasta-acertar.
— Ponderación de las alternativas del ítem.

En el método de los juicios de seguridad, además de contestar a los ítems, se les pide a las personas evaluadas que emitan un juicio acerca del grado de seguridad o confianza que tienen de acertarlo. Se asume implícitamente que a mayor seguridad en el juicio, mayor conocimiento, pero las interacciones de estos juicios con aspectos de personalidad y motivacionales de las personas son poco conocidas y, además, no se dispone de evidencia clara acerca de su influencia sobre importantes parámetros del test como fiabilidad y validez.

El método de responder-hasta-acertar consiste en eso, en que las personas van dando respuestas al ítem hasta que lo aciertan. Su implementación práctica conlleva algún artilugio que dé *feedback* a la persona evaluada para que esta sepa cuándo ha alcanzado la respuesta correcta. El instrumento ideal para desarrollar esta lógica es el ordenador, aunque también es factible con papel y lápiz. La puntuación total se obtiene penalizando el número de respuestas necesario para alcanzar las soluciones correctas. Como en el caso anterior, las ventajas respecto del modo tradicional de puntuación no resultan notables; véase al respecto Wilcox (1981, 1982).

Finalmente, el método de ponderar las alternativas consiste en asignar a cada alternativa distinto peso según su grado de corrección, estimada por expertos. El método es habitual en los ámbitos educativos, pero, como señalan Crocker y Algina (1986), hasta la fecha no hay datos concluyentes de sus beneficios.

A modo de conclusión sumaria, puede decirse que la idea general de asignar cierta calificación por el conocimiento incompleto o parcial de las respuestas es plausible, pero con los métodos desarrollados hasta ahora la evidencia empírica obtenida no es muy alentadora en lo que a ventajas sobre el sistema más clásico de todo-nada se refiere.

7. FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS

7.1. Introducción

El funcionamiento diferencial de los ítems (FDI), denominado en sus orígenes «sesgo», es un tópico que aparece tardíamente tratado, a partir de los años setenta, en la literatura psicométrica especializada. Los textos clásicos (Gulliksen, 1950; Lord y Novick, 1968; Thorndike, 1971) prácticamente ni lo citan, y tampoco lo hace la edición de 1966 de los *Standards for Educational and Psychological Tests and Manuals.* Será la discusión social y psicológica teórica, ajena en gran parte al círculo psicométrico especializado, reactivada tras el polémico artículo de Jensen (1969), la que obligue a los psicómetras profesionales a generar esta metodología, para mostrar que sus ítems y sus test no están sesgados, o sí. Además, los nuevos modelos de teoría de respuesta a los ítems proporcionarán un marco adecuado para el tratamiento del problema. Hoy la literatura al respecto es abundante. Un tratamiento enciclopédico y clásico es el de Jensen (1980), así como interesante para contrastar distintos puntos de vista en la subsiguiente revisión del libro por varios especialistas, con réplica del autor, llevada a cabo en la revista *Behavioral and Brain Sciences* (1980). El texto de Berk (1982), producto de un congreso sobre el tema, proporciona una buena visión de la tecnología disponible entonces, incluyendo además una exposición por parte de los principales editores de test norteamericanos de lo que realmente hacen para evitar el sesgo en ellos. Buenos tratamientos pueden verse en textos como los de Berk (1982), Holland y Wainer (1993), Camilli y Shepard (1994), Camilli (2006) y Osterlind y Everson (2009). En español véanse Fidalgo (1996) y Gómez, Hidalgo y Guilera (2010), donde se puede encontrar una relación de los programas informáticos más habituales para la detección del FDI. En el epígrafe 10 del capítulo 7 de este libro se expone el tratamiento del FDI dentro del marco de la teoría de respuesta a los ítems (TRI).

7.2. Concepto

Un metro estará sistemáticamente sesgado si no proporciona la misma medida para dos objetos o clases de objetos que de hecho miden lo mismo, sino que sistemáticamente perjudica a uno de ellos. Hay situaciones en las que incluso un instrumento de medida tan «sólido» como un metro puede generar medidas sesgadas. Piénsese, por ejemplo, en dos tuberías metálicas que miden 100 metros cada una; la primera conduce agua caliente a 100 grados centígrados, y la segunda, fría a 10 grados (miden lo mismo a esas temperaturas). Si se utiliza para medirlas un metro metálico, el resultado estará sesgado «contra» la tubería que conduce agua caliente, pues según este metro medirá algo menos de los 100 metros que debería. En el proceso de medir el metro metálico se habrá dilatado al calentarse y no medirá metros, sino metros y algo más, con lo cual el resultado final estará algo por debajo de 100. Puede decirse que este metro metálico está sesgado contra los tubos conductores de agua caliente, es decir, no funciona igual para ambos tubos, muestra un funcionamiento diferencial.

Un ítem mostrará FDI si para dos personas o grupos con el mismo valor en la variable medida generan mediciones distintas.

Como es fácil de entender, el problema del FDI viene acompañado de serias implicaciones sociales en el uso de los test, pues, de darse tal sesgo, ciertos grupos sociales, clásicamente blancos-negros, mujeres-hombres, pobres-ricos, etc., aunque cualquier otra partición es posible, sufrirán las consecuencias. Si se toma una postura socialmente militante y se afirma de antemano que las variables psicológicas medidas han de tomar los mismos valores para los grupos citados, u otros, entonces la definición de sesgo adoptada es mucho más lasa, a saber, se hablará de sesgo siempre que se detecten diferencias entre los grupos. Nótese que la definición original no implica esto; las comparaciones no se establecen entre los grupos considerados globalmente, sino entre las personas de ambos grupos que tienen el mismo nivel en la variable medida. Es importante entender esta diferencia, pues, aunque no muy probable, es perfectamente posible que un test esté sesgado «contra» determinada subpoblación, según el primer concepto de sesgo, y, sin embargo, este mismo grupo obtenga puntuaciones superiores a la subpoblación «favorecida» por el sesgo. Véase lo dicho en la figura 4.2, en la que se han representado en abscisas las puntuaciones globales obtenidas en el test (X) por una muestra de hombres y otra de mujeres, y en ordenadas, la proporción de aciertos de ambos grupos en un ítem (P).

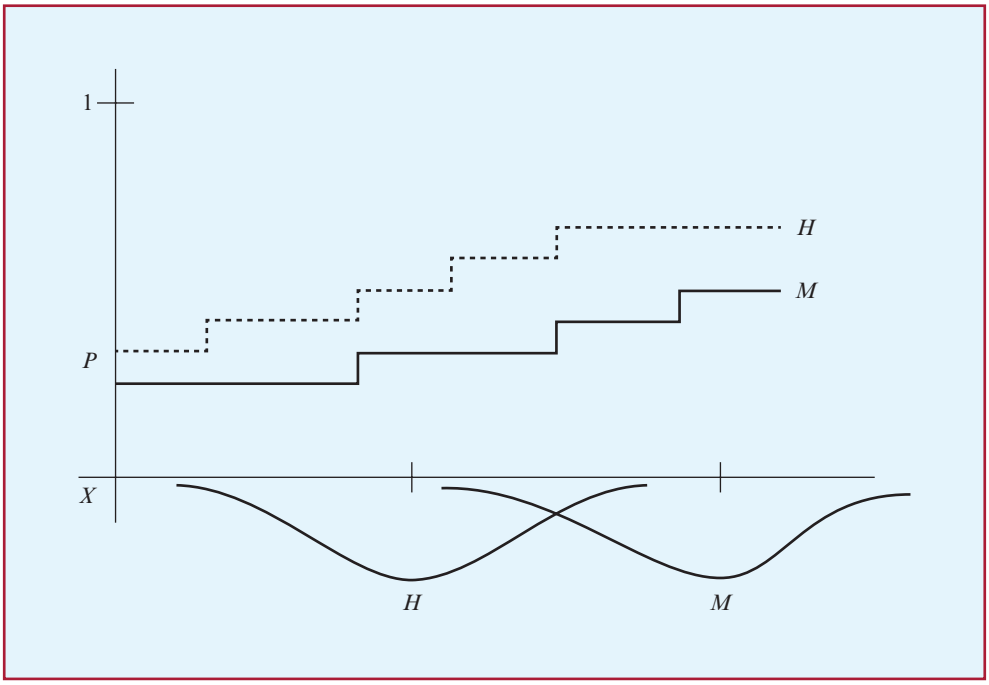


Figura 4.2—Ítem sesgado contra las mujeres.

A pesar de que el ítem está sesgado contra las mujeres para todos los valores de X (para todo valor de X la proporción de aciertos es menor en la muestra de mujeres), la media total de las mujeres en el test es superior a la de los hombres, como se puede observar en las distribuciones dibujadas en el eje de abscisas.

La psicometría se ocupa del sesgo tal como se definió en primer lugar, es decir, entiende que un ítem o un test están sesgados si personas igualmente competentes y pertenecientes a distintas subpoblaciones no tienen la misma probabilidad de superar el ítem (o test). Ahora bien, si dos personas tienen el mismo nivel en una variable, ¿a qué se puede deber que un ítem diseñado para medir esa variable pueda estar sesgado, esto es, pueda ser más favorable a uno que a otro? Las fuentes del sesgo son numerosas y vienen generadas principalmente por el distinto bagaje cultural, social, económico, etc., de las personas, o, para oídos más operantes, por la historia estimular de las personas. Dado que estos antecedentes históricos de las personas nunca serán los mismos, y pueden ser marcadamente distintos según la subcultura, si un ítem, o instrumento en general, se apoya más en la de unos que en la de otros, tendrá altísimas probabilidades de no ser equitativo, de estar sesgado. El problema puede tener serias repercusiones sociales si es precisamente una de las dos culturas, obviamente la dominante, la que construye los tests para todos. Ejemplos clásicos de sesgo se producen cuando la medición de una variable viene contaminada por otra, sesgándose la medida en función de la variable contaminadora. Si, por ejemplo, un test de competencia matemática está formulado de tal modo que exige un alto nivel de comprensión verbal, estará sesgado contra los lectores menos eficientes. En términos de diseño se confunde el efecto de la comprensión verbal con el de la competencia matemática, es decir, si una persona puntúa bajo en el test, no sabremos a ciencia cierta si atribuirlo a su bajo rendimiento en matemáticas o a que su competencia verbal es limitada y no ha llegado a captar los problemas planteados. La casuística es interminable y puede decirse que estrictamente no existen pruebas exentas completamente de sesgo; más bien se trata de detectar la cantidad de sesgo tolerable. Expuesto brevemente el concepto de sesgo, véase Shepard (1982)

para un análisis detallado. Se exponen a continuación algunas de las técnicas de las que se valen los psicómetras para la detección y análisis del sesgo.

Antes de entrar en la exposición de las técnicas para detectar el FDI, es obligado hacer una aclaración terminológica. En la actualidad prácticamente ha dejado de utilizarse el término *sesgo de los ítems* (o de los test), en favor del más aséptico *funcionamiento diferencial de los ítems* (o de los test). La razón es la siguiente. En realidad, lo que las técnicas que se van a exponer detectan es si un ítem (o un test) funciona igual o diferente para un grupo que para otro, es decir, si existe un funcionamiento diferencial del ítem (FDI) para los grupos comparados. Eso es todo, la técnica no permite ir más allá, no dice nada acerca de la naturaleza o causa del funcionamiento diferencial. Las razones del funcionamiento diferencial, si lo hubiese, corresponde buscarlas al especialista o investigador del campo. En este sentido, los resultados de aplicar los métodos estadísticos para detectar el FDI no son más que un primer paso modesto para el estudio de lo realmente importante, a saber, cuáles son las razones psicológicas, educativas, culturales, sociales, actitudinales, etc., que hacen que un ítem no funcione igual para los grupos estudiados. Por tanto, lo que se pretende con este cambio terminológico, propuesto con éxito por Holland y Thayer (1988), es una mayor precisión descriptiva de lo que realmente hacen las técnicas. Suele reservarse el término «sesgo» para el estudio más amplio que sigue una vez detectado el FDI, mediante el cual se tratan de buscar las causas que originan el funcionamiento diferencial. Nótese que de la existencia de FDI no se sigue automáticamente la existencia de sesgo, pues bien pudiera ocurrir que la causa de ese funcionamiento diferencial detectado fuera pertinente para la variable medida, con lo cual el ítem estaría cumpliendo con su cometido. Por ejemplo, imagínese un test para seleccionar controladores aéreos en el que se encuentra que algunos ítems muestran FDI, debido al distinto nivel de inglés de los aspirantes, saliendo favorecidos aquellos con mejor nivel en este idioma. Ahora bien, dado que los responsables de la selección consideran que un buen dominio del inglés es importante para desempeñar eficazmente la labor de controlador aéreo, deciden mantener estos ítems dentro del test, es decir, el funcionamiento diferencial mostrado por esos ítems no lo consideran sesgo, puesto que va en la dirección de la variable medida.

Hecha esta aclaración terminológica, en la exposición que sigue a veces se utiliza el término *sesgo* cuando en puridad se debería utilizar *funcionamiento diferencial de los ítems,* pero en cada caso el contexto permitirá al lector tener claro a qué nos estamos refiriendo.

7.3. Métodos de evaluación

Seguramente el método más eficiente para evitar en lo posible el FDI de los ítems sea un cuidadoso análisis de su contenido por parte de varios expertos previo a su publicación. Una buena exposición sobre el modo de sistematizar y formalizar esta revisión es la de Tittle (1982). Hecha tal revisión y aplicados los ítems a las personas, aún cabe llevar a cabo ciertos análisis estadísticos que permiten detectar el FDI en ítems escapados al análisis previo. A este tipo de técnicas estadísticas a posteriori nos referiremos aquí, pero dejando claro que solo son un complemento de un escrutinio riguroso previo. La nómina es abundante, pero aquí solo se abordarán cinco de las más típicas.

7.3.1. χ^2 de los aciertos

Un método para detectar el FDI derivado directamente de la definición dada consiste en dividir las puntuaciones en el test de los dos grupos estudiados en varios niveles, entre cinco y 10 normalmente, y comparar los aciertos de cada grupo en los distintos niveles. Si el ítem no está sesgado, es de esperar que las proporciones de aciertos en los distintos niveles sean iguales para los dos grupos. La significación estadística de las diferencias puede analizarse mediante la prueba de χ^2 propuesta por Scheuneman (1979) aplicada a los aciertos. Véase ilustrada en la figura 4.3 la lógica anterior para una muestra de hombres y otra de mujeres.

En el eje de abscisas aparecen representadas las puntuaciones en el test, y en el de ordenadas, la proporción de personas que acertaron el ítem considerado para cada categoría. El análisis visual del gráfico parece indicar que el ítem considerado no está sesgado, pues las proporciones de aciertos son muy

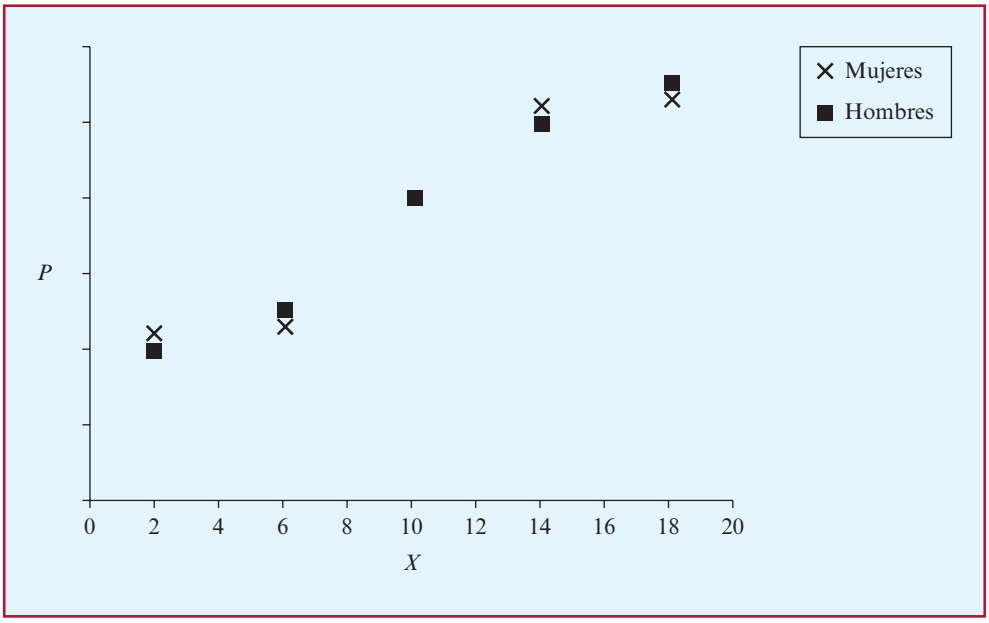


Figura 4.3.—Proporción de aciertos de un ítem en una muestra de mujeres y otra de hombres según las puntuaciones obtenidas en el test.

similares para ambos grupos. En la práctica, la situación no será probablemente tan clara y por ello se utilizan diversas técnicas estadísticas para decidir sobre la significación de las diferencias halladas. Veamos la citada de Scheuneman (1979) mediante un ejemplo.

Ejemplo

Sea un test de rapidez perceptiva aplicado a 400 personas, 200 hombres y 200 mujeres, cuyas puntuaciones se dividieron en cinco categorías según su cuantía. Se desea estudiar el posible sesgo de un ítem que por incluir estímulos perceptuales más familiares en nuestra cultura a los hombres que a las mujeres pudiera estar sesgado en contra de estas. Los resultados aparecen en la tabla 4.4.

Todos los datos de la tabla se obtienen directamente de los resultados en el test tras su corrección. Para aplicar χ^2 se necesita además la frecuencia esperada o teórica para el caso de que no existiesen diferencias entre los grupos, esto es, para el caso de

TABLA 4.4

X	Número de personas			Aciertos			Proporción aciertos (total)
	Mujeres	Hombres	Total	Mujeres	Hombres	Total	
20-24	20	15	35	15	10	25	25/35 = 0,71
15-19	100	105	205	70	85	155	155/205 = 0,76
10-14	50	40	90	10	30	40	40/90 = 0,44
5-9	20	30	50	5	20	25	25/50 = 0,50
0-4	10	10	20	0	5	5	5/20 = 0,25
	200	200	400	100	150	250	

la hipótesis nula de no sesgo. Para ello se multiplica la última columna por la primera y la segunda, obteniendo los diez valores esperados que aparecen en la tabla 4.5.

TABLA 4.5

X	Frecuencias esperadas (H_0)	
	Mujeres	**Hombres**
20-24	20 × 0,71 = 14,20	15 × 0,71 = 10,65
15-19	100 × 0,76 = 76,00	105 × 0,76 = 79,80
10-14	50 × 0,44 = 22,00	40 × 0,44 = 17,60
5-9	20 × 0,50 = 10,00	30 × 0,50 = 15,00
0-4	10 × 0,25 = 2,50	10 × 0,25 = 2,50

Los valores teóricos obtenidos de este modo son aquellos que deberían darse en caso de que la proporción de aciertos fuese la misma para ambos sexos. A partir de los valores esperados y de los obtenidos empíricamente se construye la tabla 4.6, a la que se aplica χ^2.

$$\chi^2 = \frac{(15-14{,}2)^2}{14{,}2} + \frac{(70-76)^2}{76} + \frac{(10-22)^2}{22} +$$
$$+ \frac{(5-10)^2}{10} + \frac{(0-2{,}5)^2}{2{,}5} + \frac{(10-10{,}65)^2}{10{,}65} +$$
$$+ \frac{(85-79{,}8)^2}{79{,}8} + \frac{(30-17{,}6)^2}{17{,}6} + \frac{(20-15)^2}{15} +$$
$$+ \frac{(5-2{,}5)^2}{2{,}5} = 25{,}34$$

TABLA 4.6

X	Frecuencias empíricas y teóricas			
	Mujeres		**Hombres**	
	E	***T***	***E***	***T***
20-24	15	14,20	10	10,65
15-19	70	76	85	79,80
10-14	10	22	30	17,60
5-9	5	10	20	15
0-4	0	2,50	5	2,50

El valor obtenido (25,34) es muy superior al dado por las tablas para $\chi^2_{0,99}$ con 4 grados de libertad (13,28); luego se rechaza la hipótesis nula, el ítem está sesgado. [Los grados de libertad vienen dados por $(c-1)(f-1) = (2-1)(5-1) = 4$, donde c es el número de columnas o grupos comparados, y f el de filas o niveles en los que se dividen las puntuaciones.]

Nótese que la única información proporcionada por χ^2 es que la discrepancia estadística es significativa, pero no nos indica nada acerca del sentido del sesgo, es decir, qué grupo es más favorecido por el ítem. Para ello hay que recurrir a la representación gráfica o a ciertos índices propuestos por algunos autores (Ironson, 1982; Ironson y Subkoviak, 1979) consistentes en colocar signos negativos cuando los valores empíricos de una casilla sean menores que los teóricos esperados; de ese modo, el grupo con más valores negativos sería el más perjudicado por el ítem. En nuestro ejemplo, el ítem considerado está claramente sesgado contra las mujeres, pues en cuatro de los cinco niveles los valores empíricos son menores que los teóricos, mientras que en los hombres ocurre lo contrario, como indica la figura 4.4.

La prueba de Scheuneman tiene algún inconveniente añadido a los clásicos de χ^2 de dependencia del número de sujetos y categorías, ya que estrictamente no se distribuye según χ^2, es una aproximación (Baker, 1981; Ironson, 1982; Marascuilo y Slaughter, 1981), además de no hacer uso de las respuestas incorrectas de las personas evaluadas. Este inconveniente de no usar la información proporcionada por las respuestas incorrectas de las personas puede evitarse utilizando la χ^2 global.

7.3.2. χ^2 global

Consiste en calcular χ^2 de los aciertos, tal como se ha hecho antes, y análogamente χ^2 de los errores. La suma de ambas se distribuye según la así llamada χ^2 global (Camilli, 1979) con $(K-1)j$ grados de libertad, donde K es el número de grupos comparados, y j, el número de niveles en los que se dividen las puntuaciones en el test. (Haga el lector los cálculos a modo de ejercicio para los datos del ejemplo anterior.) Una excelente comparación entre las dos χ^2, aciertos y global, amén del análisis de otros problemas implicados, como el número de categorías adecuado, número de sujetos por categoría, etc., puede verse en Ironson (1982).

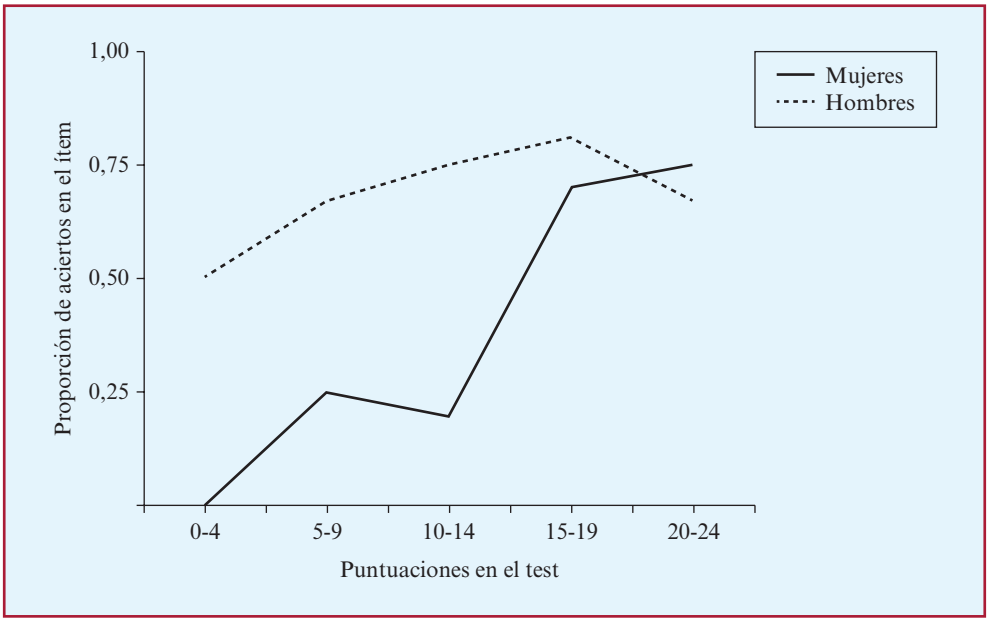


Figura 4.4.—Proporción de aciertos de un ítem en una muestra de mujeres y otra de hombres, según las puntuaciones obtenidas en el test.

El atractivo de estas técnicas basadas en χ^2 radica en su sencillez y fácil comprensión por parte de usuarios como profesores, psicólogos, médicos y otros profesionales generalmente familiarizados con χ^2. Ahora bien, aparte de los problemas de tipo, digamos, técnico-estadístico apuntados, su debilidad radica en que se apoyan en la asunción de que la mayoría de los ítems del test no están sesgados. Nótese que al establecer las categorías en las puntuaciones del test para evaluar el sesgo de un ítem se asume que los $(n - 1)$ ítems utilizados para obtenerlas reflejan las verdaderas puntuaciones de los dos grupos a comparar, esto es, son insesgados. Esto no hay medio de comprobarlo; luego si ocurriera que la mayoría de los ítems estuvieran sesgados, así lo estarían las puntuaciones de las personas en cada nivel. En suma, las técnicas anteriores serían útiles si la asunción de que la mayoría de los ítems están insesgados se cumple, asunción razonable pero difícilmente contrastable empíricamente; en una situación de sesgo generalizado constituyen un razonamiento circular, o, más exactamente, solo permitirían comprobar el sesgo de un ítem respecto del sesgo de otros ítems que conforman el test tomados conjuntamente.

Una forma de mitigar el problema es proceder por etapas. En una primera fase se detectan mediante alguno de los métodos los ítems que presentan FDI. En una segunda etapa se reanaliza el funcionamiento diferencial de los ítems, utilizando para establecer las categorías únicamente los ítems que no presentaron FDI en la primera fase. De este modo se purifica notablemente la puntuación global a partir de la cual se establecen las categorías. Abundantes investigaciones muestran que estos procedimientos iterativos funcionan bastante mejor que los realizados en una sola etapa.

7.3.3. Método *delta*

El método *delta* (Angoff y Ford, 1973; Angoff, 1982b) ha sido uno de los más utilizados antes del desarrollo de la metodología actual. A grandes rasgos, consiste en calcular las proporciones de aciertos o índices de dificultad clásicos de cada ítem para los grupos en los que se pretende estudiar el FDI. Estas proporciones se convierten en puntuaciones típicas bajo la curva normal, puntuaciones que se transforman a su vez en otra escala más ma-

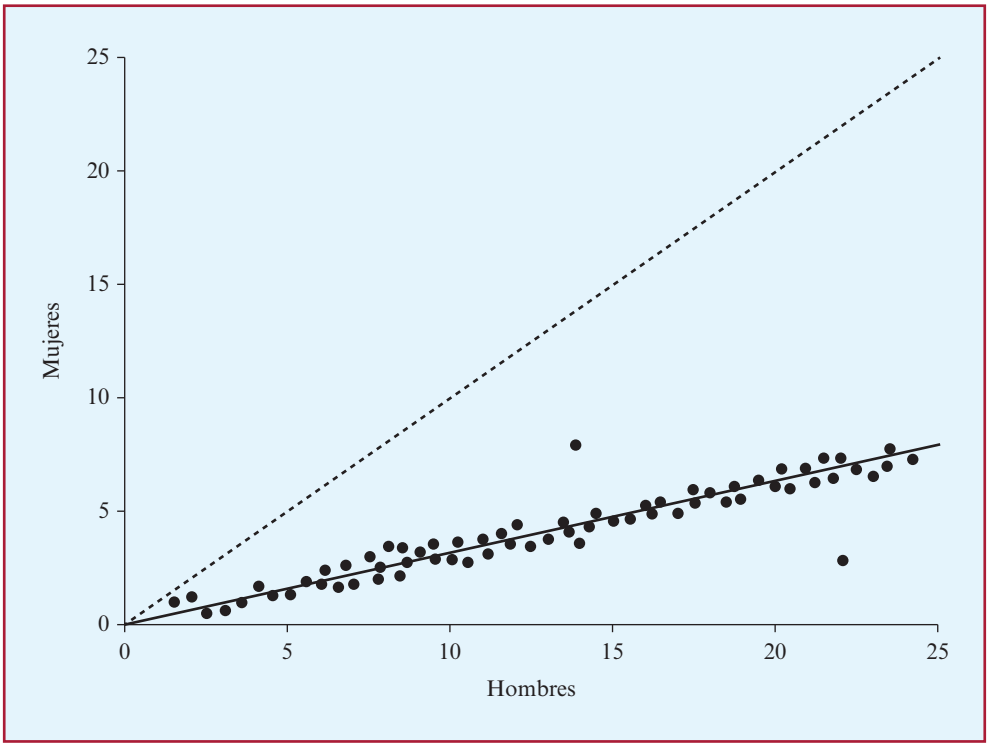


Figura 4.5.—Puntuaciones *delta.*

nejable. La más afamada y clásica es la escala *delta* y, en consecuencia, las puntuaciones *delta,* obtenidas al multiplicar las típicas por 4 y sumarles 13. Una vez halladas las puntuaciones *delta* para cada grupo, se representan gráficamente. Si todos los puntos caen en una recta, los ítems estarían insesgados, considerándose sesgados aquellos que se apartan sustancialmente de la recta. Véase ello ilustrado en la figura 4.5, donde según este criterio dos de los ítems estarían especialmente sesgados.

Nótese que el test en cuestión es más fácil para las mujeres que para los hombres, al estar los valores de los ítems por debajo de la diagonal; la dificultad sería la misma para ambas subpoblaciones cuando los valores se ajustasen a la diagonal. No confundir dificultad y sesgo.

Un índice general del FDI viene dado por el ajuste de los puntos a la recta, lo que se puede hallar mediante la correlación de Pearson entre los valores *delta* de los ítems para los dos grupos comparados. Angoff y Ford (1973) proponen también como indicador del FDI de cada ítem su distancia al eje principal de la elipse generada por los ítems. Cuanto mayor sea la distancia, más FDI mostrará el ítem respecto de los otros.

El eje principal viene dado por:

$$Y = aX + b$$

donde

$$a = \frac{S_y^2 - S_x^2 + \sqrt{(S_y^2 - S_x^2)^2 + 4r_{xy}^2 S_x^2 S_y^2}}{2r_{xy} S_x S_y}$$

$$b = \overline{Y} - a\overline{X}$$

y el índice de distancia por:

$$d = \frac{aX_j - Y_j + b}{\sqrt{a^2 + 1}}$$

donde

X_j: Puntuación *delta* del ítem *j* en el grupo *X*.
Y_j: Puntuación *delta* del ítem *j* en el grupo *Y*.
a y *b*: Vienen dados por las fórmulas citadas.

Calculada la distancia para todos los ítems, su distribución proporciona una idea de conjunto del FDI y permite establecer criterios de corte descriptivos para descartar ítems que se apartan notablemente de la media.

Conviene hacer algunas puntualizaciones sobre lo dicho. La razón de la transformación de los índices de dificultad o proporciones (*P*) de aciertos de los ítems en puntuaciones típicas (*Z*) tiene como objetivo convertir en lineales las relaciones que entre las proporciones directamente podrían ser curvilíneas. Dicha transformación se denomina «inversa normal», porque asigna no a *P* sino a (1 − *P*) la *Z* correspondiente. La legitimidad teórica de esta transformación es cuestionada por Lord (1980), y en consecuencia el método mismo. Nótese también que si todos los puntos generados por los ítems se ajustan adecuadamente a una recta, según el criterio anterior no existiría FDI, pero es igualmente lícito afirmar que todos los ítems están igual de sesgados. Más que el FDI en abstracto, el método *delta* evalúa la discrepancia entre sesgos. De nuevo, como ocurría con χ^2, se precisa la asunción de que la mayoría de los ítems no presentan FDI.

7.3.4. Mantel-Haenszel

Debido a su relativa sencillez de cálculo y a los buenos resultados que ofrece, el método de Mantel y Haenszel (1959) es el más utilizado en la actualidad. Su aplicación a la evaluación del funcionamiento diferencial de los ítems fue propuesta inicialmente por Holland (Holland, 1985; Holland y Thayer, 1985, 1986, 1988) y puede considerarse una extensión natural de los métodos de χ^2 expuestos previamente.

La lógica general del método es clara y sencilla: un ítem no presentará un funcionamiento diferencial si el cociente entre las personas que aciertan el ítem y las que lo fallan es el mismo para los dos grupos comparados en cada una de las categorías o niveles en los que se dividan las puntuaciones del test. Esa es la hipótesis nula que se somete a prueba mediante el estadístico de Mantel-Haenszel (M-H). Formalmente esta hipótesis puede formularse del siguiente modo:

$$H_0: \frac{A_j}{B_j} = \frac{C_j}{D_j} \text{ para cada una de las categorías } j$$

donde *A, B, C* y *D* son las frecuencias absolutas correspondientes a cada una de las categorías *j* en las que se dividen las puntuaciones del test, según se indica en la tabla 4.7 adjunta.

TABLA 4.7

	Aciertos (1)	Errores (0)	Marginales
Grupo de referencia (*R*)	A_j	B_j	n_{Rj}
Grupo focal (*F*)	C_j	D_j	n_{Fj}
Marginales	n_{1j}	n_{0j}	N_j

Para aplicar el método de M-H a unos datos se procede igual que se hacía para la prueba de χ^2 de los aciertos. Lo primero que hay que hacer es dividir la muestra en varias categorías o intervalos en función de las puntuaciones globales del test, y posteriormente computar los aciertos en el ítem cuyo funcionamiento diferencial se indaga para cada una de las categorías y grupos. En suma, la estructura de datos de la que se parte es la misma que la utilizada en la prueba χ^2 de los aciertos. A partir de ahí, se construye una tabla como la 4.7 de dos filas y dos columnas (2 × 2) para cada una de las categorías.

Es arbitrario asignar a un grupo u otro de los que se estudia la denominación de referencia o focal, aunque suele reservarse el término «focal» para aquel grupo, generalmente minoritario, que a priori se considera posiblemente perjudicado por alguno de los ítems. No obstante, esto no es esencial para aplicar el método, siempre y cuando seamos consecuentes con la asignación hecha a la hora de interpretar los resultados.

El estadístico de Mantel-Haenszel viene dado por la siguiente fórmula:

$$\chi^2_{MH} = \frac{\left(\left|\sum_j A_j - \sum_j E(A_j)\right| - 0{,}5\right)^2}{\sum_j \mathrm{Var}(A_j)} \quad [4.16]$$

donde

χ^2_{MH}: Se distribuye según χ^2 con 1 grado de libertad.

$\Sigma_j A_j$: Representa la suma de los valores de A para cada una de las categorías j.

$\Sigma_j E(A_j)$: Es la suma de las esperanzas matemáticas de A, que para cada una de las categorías j viene dada por:

$$E(A_j) = \frac{n_{Rj} n_{1j}}{N_j}$$

$\Sigma_j \text{Var}(A_j)$: Es la suma de las varianzas de A para cada una de las categorías j, que viene dada por:

$$\text{Var}(A_j) = \frac{n_{Rj} n_{Fj} n_{1j} n_{0j}}{N_j^2 (N_j - 1)}$$

Veamos paso a paso cómo se lleva a cabo el cálculo. Para ello utilizaremos los datos de la tabla 4.4. Recuérdese que la tabla refleja los resultados de una muestra de 400 personas (200 hombres y 200 mujeres) tras responder a uno de los ítems de un test de rapidez perceptiva que se sospechaba podría estar sesgado en contra de las mujeres.

En primer lugar, hay que confeccionar tantas tablas de dos filas y dos columnas (2 × 2) como intervalos se hayan formado a partir de las puntuaciones globales del test. En nuestro ejemplo serán cinco tablas, pues se han establecido cinco intervalos. Debajo de cada tabla aparecen los cálculos que se utilizarán posteriormente para hallar el estadístico de Mantel-Haenszel.

Intervalo 0-4

	Aciertos	Errores	Marginales
Hombres	5	5	10
Mujeres	0	10	10
Marginales	5	15	20

$$A_1 = 5$$

$$E(A_1) = \frac{(5 \times 10)}{20} = 2{,}5$$

$$\text{Var}(A_1) = \frac{(10 \times 10 \times 5 \times 15)}{20^2 (20 - 1)} = 0{,}987$$

Intervalo 5-9

	Aciertos	Errores	Marginales
Hombres	20	10	30
Mujeres	5	15	20
Marginales	25	25	50

$$A_2 = 20$$

$$E(A_2) = \frac{(25 \times 30)}{50} = 15$$

$$\text{Var}(A_2) = \frac{(30 \times 20 \times 25 \times 25)}{50^2 (50 - 1)} = 3{,}061$$

Intervalo 10-14

	Aciertos	Errores	Marginales
Hombres	30	10	40
Mujeres	10	40	50
Marginales	40	50	90

$$A_3 = 30$$

$$E(A_3) = \frac{(40 \times 40)}{90} = 17{,}777$$

$$\text{Var}(A_3) = \frac{(40 \times 50 \times 40 \times 50)}{90^2 (90 - 1)} = 5{,}549$$

Intervalo 15-19

	Aciertos	Errores	Marginales
Hombres	85	20	105
Mujeres	70	30	100
Marginales	155	50	205

$$A_4 = 85$$

$$E(A_4) = \frac{(155 \times 105)}{205} = 79{,}39$$

$$\text{Var}(A_4) = \frac{(105 \times 100 \times 155 \times 50)}{205^2 (205 - 1)} = 9{,}492$$

Intervalo 20-24

	Aciertos	Errores	Marginales
Hombres	10	5	15
Mujeres	15	5	20
Marginales	25	10	35

$$A_5 = 10$$

$$E(A_5) = \frac{(25 \times 15)}{35} = 10{,}714$$

$$\text{Var}(A_5) = \frac{(15 \times 20 \times 25 \times 10)}{35^2(35-1)} = 1{,}8$$

Una vez confeccionadas las tablas de contingencia de 2 × 2 para cada una de las cinco categorías, y obtenidos los valores de A, $E(A)$ y Var (A) para cada una de ellas, se obtienen las sumas correspondientes y se aplica la fórmula 4.16.

$$\sum_j A_j = 5+20+30+85+10 = 150$$

$$\sum_j E(A_j) = 2{,}5+15+17{,}777+79{,}39+10{,}714 = 125{,}381$$

$$\sum_j \text{Var}(A_j) = 0{,}987+3{,}061+5{,}549+9{,}492+1{,}8 = 20{,}889$$

$$\chi^2_{MH} = \frac{(|150-125{,}381|-0{,}5)^2}{20{,}889} = 27{,}85$$

Al nivel de confianza del 99% con 1 grado de libertad el valor de χ^2 en las tablas viene dado por 6,63. Puesto que 27,85 > 6,63, la diferencia entre el grupo de referencia (hombres) y el focal (mujeres) resulta estadísticamente significativa; por tanto el ítem analizado no funciona igual para los hombres y las mujeres. Este resultado coincide con el obtenido en el apartado 7.3.1 para los mismos datos por el método de χ^2 de los aciertos. En este caso el funcionamiento diferencial del ítem es tan obvio que todos los métodos lo detectan sin problema; es a medida que las diferencias entre los grupos se van haciendo más pequeñas cuando pueden surgir algunas discrepancias entre los distintos métodos utilizados.

Fijado un cierto nivel de confianza, el método de Mantel-Haenszel solo indica si el ítem funciona diferencialmente o no para los grupos estudiados, pero no informa ni sobre el grupo perjudicado por el funcionamiento diferencial del ítem ni sobre la cuantía de las diferencias en funcionamiento. La forma más sencilla de averiguar estas dos cuestiones es representando gráficamente las proporciones de aciertos de cada grupo para las distintas categorías formadas, como se indica en la figura 4.4.

Mantel-Haenszel proporcionan un estimador numérico que indica la cuantía y dirección de las diferencias de funcionamiento encontradas. El estimador viene dado por:

$$\hat{\alpha}_{MH} = \frac{\sum_j \frac{A_j D_j}{N_j}}{\sum_j \frac{B_j C_j}{N_j}} \qquad [4.17]$$

Para aplicarlo a los datos del ejemplo anterior hay que hacer los cálculos para cada una de las cinco categorías establecidas y luego sumarlos, como indica la fórmula.

$$\frac{A_1 D_1}{N_1} = \frac{5 \times 10}{20} = 2{,}5$$

$$\frac{B_1 C_1}{N_1} = \frac{5 \times 0}{20} = 0$$

$$\frac{A_2 D_2}{N_2} = \frac{20 \times 15}{50} = 6$$

$$\frac{B_2 C_2}{N_2} = \frac{10 \times 5}{50} = 1$$

$$\frac{A_3 D_3}{N_3} = \frac{30 \times 40}{90} = 13{,}33$$

$$\frac{B_3 C_3}{N_3} = \frac{10 \times 10}{90} = 1{,}11$$

$$\frac{A_4 D_4}{N_4} = \frac{85 \times 30}{205} = 12{,}44$$

$$\frac{B_4C_4}{N_4} = \frac{20 \times 70}{205} = 6{,}83$$

$$\frac{A_5D_5}{N_5} = \frac{10 \times 5}{35} = 1{,}43$$

$$\frac{B_5C_5}{N_5} = \frac{5 \times 15}{35} = 2{,}14$$

$$\hat{\alpha}_{MH} = \frac{(2{,}5 + 6 + 13{,}33 + 12{,}44 + 1{,}43)}{(0 + 1 + 1{,}11 + 6{,}83 + 2{,}14)} = 3{,}22$$

Los valores de $\hat{\alpha}_{MH}$ oscilan entre cero e infinito. Valores mayores que 1 indican que el ítem favorece al grupo de referencia, y menores, al focal. En consecuencia, para nuestro ejemplo el ítem analizado favorece notablemente a los hombres, dado que el valor obtenido (3,22) se aleja claramente de 1.

Una sencilla transformación propuesta por Holland y Thayer (1985) permite expresar el valor de $\hat{\alpha}_{MH}$ en una escala simétrica con origen cero, lo cual resulta muy útil cara a la interpretación de los resultados. A medida que los valores se alejan de cero, aumenta el funcionamiento diferencial, indicando los valores negativos que el ítem beneficia al grupo de referencia, y los positivos al focal.

La transformación viene dada por:

$$\Delta_{MH} = -2{,}35 \ln (\hat{\alpha}_{MH}) \qquad [4.18]$$

donde Δ_{MH} es la nueva métrica y ln el logaritmo neperiano de $(\hat{\alpha}_{MH})$.

Para los datos de nuestro ejemplo:

$$\Delta_{MH} = -2{,}35 \ln (3{,}22) = -2{,}75$$

Este resultado pone de manifiesto que el ítem tiene un claro funcionamiento diferencial para ambos grupos, dado que el valor se aleja notablemente de cero, y que favorece al grupo de referencia, puesto que el signo es negativo.

Como señalan Holland y Thayer (1985), esta escala de Δ_{MH} puede interpretarse también como una medida del funcionamiento diferencial de los ítems en la escala de diferencias de dificultad entre los ítems expresadas por la escala delta expuesta en el apartado anterior.

Una pregunta que tal vez le surja al lector es el número de categorías o intervalos que deben hacerse. Lo ideal es hacer tantas como ítems tiene el test más una. Lo que ocurre es que si uno de los grupos a comparar, generalmente el focal, está formado por pocas personas, puede ocurrir que muchas de estas categorías queden vacías, por lo que hay que agruparlas. A medida que se reduce el número de categorías, tiende a aumentar la probabilidad de catalogar ítems con funcionamiento diferencial cuando en realidad no lo tienen. Tiende, en suma, a aumentar el error de tipo I, o falsos positivos, sobre todo si hay diferencias en el test entre los grupos comparados. No existe un número mágico de categorías ni de personas por categoría, en cada caso hay que llegar a un compromiso entre el número de categorías y el número mínimo de personas por categoría. En general, como señalan Hambleton, Clauser, Mazor y Jones (1993), no conviene utilizar el método de Mantel-Haenszel cuando uno de los grupos (focal o referencia) tiene menos de 200 personas. A veces los educadores y psicólogos no disponen de ese número mínimo de personas, en cuyo caso pueden al menos recurrir a representaciones gráficas, pues la mera inspección visual introduce notables ganancias en relación con no hacer nada.

Al estar tan extendido el uso del método de Mantel-Haenszel, se han realizado numerosos trabajos de investigación acerca de su funcionamiento bajo todo tipo de circunstancias, un resumen de los cuales puede consultarse en Hambleton, Clauser, Mazor y Jones (1994) o Fidalgo (1996). La limitación más reseñable del método es que no detecta cuando existe lo que se llama «funcionamiento diferencial no uniforme», si bien este tipo de FDI no suele ser muy habitual en la práctica. Se dice que existe FDI uniforme cuando el ítem perjudica sistemáticamente a uno de los grupos a lo largo de todas las categorías en las que se dividen las puntuaciones del test. Por el contrario, cuando para unas categorías el ítem perjudica a uno de los grupos y para otras perjudica al otro grupo, se dice que tiene un FDI no uniforme. Por ejemplo, en las figuras 4.3 y 4.4 puede observarse un FDI no uniforme para hombres y mujeres; nótese que se cruzan las líneas que unen las proporciones de aciertos de ambos grupos para los distintos niveles. El análisis de las causas de ese tipo de funcionamiento no uniforme resulta de sumo interés para investigadores y profesionales. Para solucionar esta limitación del método de M-H, Mazor, Clauser y Hambleton (1994) han

propuesto con éxito una variante sencilla, consistente en dividir la muestra en dos grupos (por encima y por debajo de la media total) y hacer los cálculos por separado para cada uno de los grupos.

7.3.5. Índice de estandarización

Se expone finalmente un indicador del funcionamiento diferencial de los ítems de uso muy extendido. Su cálculo exige, como los anteriores, dividir en varias categorías a las personas de la muestra en función de sus puntuaciones en el test. El índice de estandarización (Dorans y Holland, 1993) cuantifica las diferencias entre las proporciones de aciertos de los grupos de referencia y focal para cada una de las categorías *j* en las que se divide el test, ofreciendo un indicador global de esas diferencias para el ítem.

El índice de estandarización viene dado por la siguiente fórmula:

$$IE = \frac{\sum_j n_{Fj}(P_{Fj} - P_{Rj})}{\sum_j n_{Fj}} \qquad [4.19]$$

donde:

- n_{Fj}: Número de personas del grupo focal para cada una de las categorías *j* (véase la tabla 4.7).
- P_{Fj}: Proporción de personas del grupo focal que aciertan el ítem para la categoría *j*. En términos de la tabla 4.7:

$$P_{Fj} = \frac{C_j}{n_{Fj}}$$

- P_{Rj}: Proporción de personas del grupo de referencia que aciertan el ítem para la categoría *j*. En términos de la tabla 4.7:

$$P_{Rj} = \frac{A_j}{n_{Rj}}$$

El índice de estandarización (IE) varía entre −1 y +1. Los valores positivos indican que el ítem beneficia al grupo focal, y los negativos, al de referencia. Dorans y Holland (1993) proponen una serie de valores que resultan muy útiles para la interpretación práctica de los resultados obtenidos al aplicar el IE:

- Entre −0,05 y 0,05 no existiría funcionamiento diferencial del ítem.
- Entre −0,05 y −0,10, o 0,05 y 0,10, conviene inspeccionar el ítem.
- Valores fuera del intervalo (−0,10; 0,10) exigen examinar el ítem concienzudamente, pues el funcionamiento diferencial es claro.

Veamos la aplicación de la fórmula 4.19 a los datos del ejemplo utilizado para ilustrar los cálculos del método de Mantel-Haenszel. Para ello se calcula el valor del numerador para cada una de las cinco tablas correspondientes a los cinco intervalos establecidos entonces.

$$n_{F1}(P_{F1} - P_{R1}) = 10\left(\frac{0}{10} - \frac{5}{10}\right) = -5$$

$$n_{F2}(P_{F2} - P_{R2}) = 20\left(\frac{5}{20} - \frac{20}{30}\right) = -8{,}33$$

$$n_{F3}(P_{F3} - P_{R3}) = 50\left(\frac{10}{50} - \frac{30}{40}\right) = -27{,}50$$

$$n_{F4}(P_{F4} - P_{R4}) = 100\left(\frac{70}{100} - \frac{85}{105}\right) = -10{,}95$$

$$n_{F5}(P_{F5} - P_{R5}) = 20\left(\frac{15}{20} - \frac{10}{15}\right) = 1{,}67$$

A partir de estos datos ya se puede aplicar la fórmula 4.19:

$$IE = \frac{[(-5) + (-8{,}33) + (-27{,}5) + (-10{,}95) + 1{,}67]}{(10 + 20 + 50 + 100 + 20)} =$$

$$= -0{,}25$$

Dado que el valor obtenido (−0,25) excede con mucho −0,10, el ítem presenta un claro funciona-

miento diferencial, favoreciendo al grupo de referencia, puesto que el signo es negativo. Este resultado coincide con los encontrados para los mismos datos mediante los métodos de Mantel-Haenszel y χ^2 de los aciertos.

Dorans y Holland (1993) ofrecen una transformación del IE a la métrica delta, así como un error típico de medida que permite someter a prueba la significación estadística de las diferencias encontradas entre los grupos de referencia y focal. Como se puede obsevar en la fórmula 4.19, el índice de estandarización utiliza el número de personas del grupo focal en cada categoría (n_{Fj}) para ponderar las diferencias entre las proporciones de aciertos de ambos grupos ($P_{Fj} - P_{Rj}$). Esta ponderación es muy razonable, puesto que permite dar más peso a aquellas categorías que cuentan con mayor número de personas del grupo focal. No obstante, Dorans y Holland (1993) proponen otras posibles ponderaciones, en función de los intereses específicos de los investigadores.

7.3.6. Comentarios finales

Hay que señalar que sigue siendo muy importante para los usuarios la representación gráfica de los datos (figuras 4.3 y 4.4), pues aunque los diversos métodos ayuden a tomar decisiones con fundamento estadístico, la visualización de las proporciones de aciertos de los grupos de referencia y focal para cada una de las categorías permite detectar el tipo de FDI existente y tener una primera aproximación de su cuantía. Los gráficos, si además se representan en abscisas las distribuciones de las puntuaciones del test para los dos grupos comparados, permiten distinguir entre dos conceptos bien distintos y a menudo confundidos: el funcionamiento diferencial del ítem y las posibles diferencias reales de los dos grupos en el ítem, lo que se denomina «el impacto». Nótese que puede haber claras diferencias en el número de aciertos de cada uno de los grupos en el ítem (impacto), y sin embargo no existe FDI, como se ilustra, por ejemplo, en la figura 4.6. Obsérvese cómo la media de aciertos en el ítem para el grupo de mujeres (0,80) es mucho más elevada que la de los hombres (0,60), por lo que existe impacto. Sin embargo, las proporciones de aciertos por categorías coinciden, los puntos y las aspas se superponen para todas las categorías; por tanto, no hay FDI. Globalmente el ítem resulta más difícil para los hombres, pero no se puede decir que funcione diferencialmente para los hombres y las mujeres; en suma, hay impacto pero no FDI. La distinción es clave, pues entre dos grupos nada impide que haya diferencias (impacto), en ningún lugar está escrito que todos los grupos deban obtener las mismas puntuaciones en todas las variables, pero lo que hay que evitar a toda costa es un funcionamiento diferencial de los ítems para los grupos. La confusión de estos dos conceptos es el origen de muchas de las necedades que sobre el sesgo de los test se han dicho. Un caso ya clásico es el acaecido al Educational Testing Service (ETS), una de las empresas editoras y aplicadoras de test más potentes de los Estados Unidos, que fue demandada en los tribunales por una compañía de seguros (Golden Rule Insurance Company) alegando esta que algunos de los test utilizados por el ETS estaban sesgados contra los negros. Tras ocho años peleando en los tribunales, en 1984 el presidente del ETS, intentando quitarse el engorroso asunto de encima, muy costoso en tiempo y dinero, llegó a un acuerdo con la compañía de seguros para descartar en el futuro todos aquellos ítems en los que la proporción de aciertos de los blancos excediesen en 0,15 a los de los negros. El presidente sabía de sobra que esto era injustificado, pues los ítems han de descartarse cuando muestran FDI, no cuando hay impacto, pero seguramente consideró que sería una forma práctica de zanjar la cuestión. Todo lo contrario, la polvareda levantada por parte de los expertos fue enorme, y el hombre tuvo que dar marcha atrás, admitiendo que había cometido un error. El lector interesado en los intríngulis de este asunto puede consultar el número especial dedicado al tema por la revista *Educational Measurement: Issues and Practice* (1987), donde se analiza desde diferentes perspectivas.

Todas las técnicas expuestas previamente suelen catalogarse como *internas,* en alusión a que el criterio de contraste para analizar los ítems es interno al test; recuérdese que para establecer las distintas categorías se utilizaba la puntuación global de las personas en el test o, en el caso de proceder en dos etapas, la puntuación en el conjunto de ítems que no presentaban FDI en un primer análisis. Por el contrario, se habla de FDI *externo* cuando el criterio de contraste es externo, por ejemplo el criterio

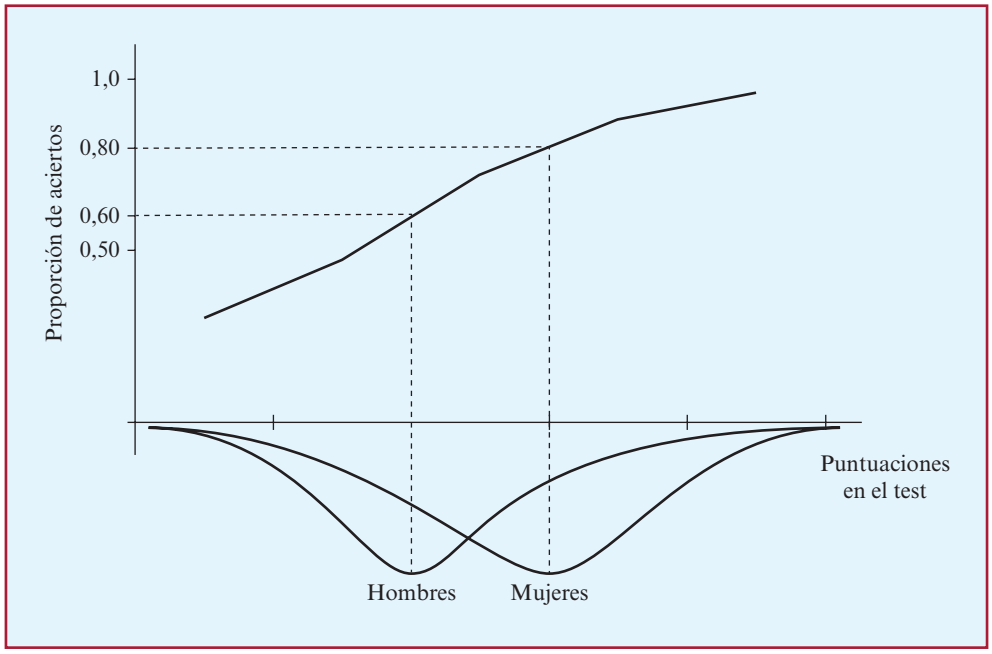


Figura 4.6.—Impacto y funcionamiento diferencial de un ítem para un grupo de hombres y otro de mujeres.

que se pretende predecir con el test. La estrategia más habitual en este caso es calcular la recta de regresión del criterio externo sobre el test para la muestra total y para cada uno de los grupos (focal y referencia) y compararlas. Si no existiese FDI, las tres rectas deberían coincidir. No se han expuesto las técnicas externas, puesto que en la actualidad son poco utilizadas.

Señalar finalmente que las técnicas expuestas son todas ellas de carácter *condicional,* excepto el método *delta*, que sería *incondicional.* En las técnicas condicionales los aciertos en el ítem estudiado se contrastan condicionalmente para cada una de las categorías formadas, mientras que en las incondicionales, como el método *delta,* no se establecen categorías, se utilizan las puntuaciones globales de los grupos focal y de referencia. Esta incondicionalidad es precisamente lo que hace delicado el uso del método *delta,* así como otras técnicas incondicionales, pues pueden llevar a confundir funcionamiento diferencial con impacto.

El funcionamiento diferencial de los ítems es uno de los campos en los que más se ha desarrollado la teoría de los test estos últimos años, habiéndose propuesto otros muchos métodos que no exponemos aquí, dado el carácter introductorio de este libro. Cabría destacar el SIBTEST (Shealy y Stout, 1993), las técnicas para analizar las tablas de contingencia (modelos loglineales, modelos logit, regresión logística) o toda la tecnología basada en la teoría de respuesta a los ítems. El lector interesado puede consultar el libro de Camilli y Shepard (1994), o en castellano el trabajo de Fidalgo (1996), y para una buena revisión, Gómez, Hidalgo y Guilera (2010). En el epígrafe 10 del capítulo 7 se presenta la estimación del FDI en el marco de la TRI.

EJERCICIOS

1. En la tabla adjunta aparecen las puntuaciones obtenidas por una muestra de cinco sujetos en un test de cuatro ítems, así como sus puntuaciones en el criterio.

Sujetos	ítems				Criterio
	1	2	3	4	
A	1	1	0	1	5
B	0	1	1	0	2
C	1	1	1	0	4
D	1	0	0	0	1
E	0	0	0	0	1

1. Calcule el índice de dificultad de todos los ítems.
2. Calcule el índice de discriminación del ítem 4.
3. Calcule el índice de validez del ítem 3.
4. Calcule el coeficiente de validez del test.

2. Una muestra de seis alumnos universitarios obtuvo las siguientes puntuaciones en un test de coordinación visomotora:

Sujetos	Test	Criterio
A	2	2
(B)	4	0
C	4	3
(D)	5	3
E	6	4
(F)	9	6

Únicamente los tres que aparecen entre paréntesis acertaron el último ítem del test, aunque todos lo intentaron resolver.

1. Calcule el índice de dificultad del último ítem.
2. Calcule el índice de discriminación del último ítem.
3. Calcule el índice de validez del último ítem.
4. Calcule el número de discriminaciones que realiza entre los sujetos el último ítem.
5. Calcule el coeficiente de validez del test.

3. Un test consta de 150 ítems de tres alternativas, una de las cuales es correcta.

1. Un sujeto que desconozca por completo el contenido del test, pero aun así responda a todos los ítems, ¿cuántos es de esperar que acierte?
2. ¿Qué puntuación directa se le asignará una vez corregidos los efectos del azar?

4. A una muestra de 500 estudiantes universitarios se le aplicó un test de 260 ítems. Los 80 primeros solo tenían dos alternativas, una de ellas correcta. Los 120 siguientes constaban de cinco alternativas, una correcta, y los 60 restantes tenían cuatro alternativas, también con solo una correcta.

1. Uno de los sujetos responde a todos los ítems, de los 80 primeros acierta 70; de los 120 siguientes falla 20, y de los restantes acierta tantos como falla.
   ¿Qué puntuación directa se le asignará una vez corregidos los efectos del azar?
2. Si un sujeto no sabe nada y contesta al azar, ¿cuántos ítems cabe esperar que acierte? ¿Qué puntuación directa se le asignará tras la corrección de los efectos del azar?

5. En la matriz adjunta aparecen las puntuaciones obtenidas por cinco sujetos en un test de cuatro ítems, así como los índices de validez de los ítems.

Sujetos	Ítems			
	1	2	3	4
A	1	1	0	1
B	0	1	1	0
C	1	1	1	0
D	1	0	0	0
E	0	0	0	0
Índices de validez	0,1	0,2	0,6	0,25

1. Calcular los índices de discriminación de los ítems.
2. Calcular el coeficiente de validez del test.

6. En la tabla adjunta aparecen las respuestas dadas por 200 sujetos a las cinco alternativas (*A*, *B*, *C*, *D*, *E*) del ítem 19 de un test de 40, y de las que solo la *C* es correcta. Se han separado por un lado las respuestas dadas por los sujetos que obtuvieron puntuaciones superiores a la mediana del test, y por otro aquellos sujetos con puntuaciones inferiores a la mediana. También se exponen en la parte inferior de la tabla las medias obtenidas en el test por los

sujetos que respondieron a cada alternativa. La desviación típica de las puntuaciones de los sujetos en el test fue 10.

	A	B	C	D	E
50% superior	5	15	60	18	2
50% inferior	40	15	20	15	10
Media test	15	20	24	18	14

1. Calcular el índice de dificultad del ítem.
2. Analizar las respuestas de los sujetos a las alternativas incorrectas del ítem y señalar aquellas que no contribuyen a discriminar entre los sujetos competentes y los incompetentes en el test. Razonar adecuadamente.
3. Calcular el índice de discriminación del ítem.
4. Calcular la covarianza entre el ítem y el test.

SOLUCIONES

1.1. 0,6; 0,6; 0,4; 0,2
2. 0,19
3. 0,20
4. 0,907

2.1. 0,5
2. 0,25
3. 0,00
4. 9
5. 0,80

3.1. 50
2. 0

4.1. 175
2. 79; 0

5.1. 0,16; 0,77; 0,08; 0,19
2. 0,95

6.1. 0,25
2. *B*, *D*
3. 0,34
4. 1,66

5 Transformación de las puntuaciones

Una vez que se han obtenido las puntuaciones de las personas en un test, para facilitar su interpretación y comprensión por parte de los interlocutores y clientes, las puntuaciones directas suelen transformarse en otros tipos de puntuaciones. Como ahora se verá, varias transformaciones son posibles. El objetivo fundamental de las transformaciones es expresar las puntuaciones directas de tal modo que hagan alusión a la ubicación de la persona en el grupo, dando así la idea comparativa de su puntuación en relación con sus semejantes. Por ejemplo, si tras realizar un test se nos dice que hemos obtenido en él 80 puntos, no tenemos ni idea de lo que eso representa respecto a nuestros colegas; ¿estamos por encima de la media?, ¿por debajo?, etc. Pues bien, ese es el tipo de información que pretenden dar las puntuaciones directas una vez transformadas: la ubicación relativa en el grupo, constituyendo un indicador del escalamiento de las personas. Nótese que las puntuaciones transformadas no añaden ninguna información a la contenida en las directas, salvo las ventajas prácticas que pueda tener esta forma de ofrecer la información. Ni que decir tiene que las puntuaciones transformadas no alteran el escalamiento hecho por las directas, sencillamente lo expresan en otra escala. Para los análisis estadísticos y psicométricos deben utilizarse siempre las puntuaciones directas obtenidas y no las transformadas.

1. PERCENTILES

La transformación a escala centil consiste en asignar a cada puntuación directa el porcentaje de personas que obtienen puntuaciones inferiores a ella. Dan, por tanto, una idea rápida e intuitiva de la posición relativa de la persona en el grupo. Así, por ejemplo, una persona con un percentil de 80 indicaría que su puntuación en el test es superior al 80% de sus compañeros. Esta transformación es, sin duda, la más utilizada, debido sobre todo a su simplicidad y universalidad, lo que facilita la interacción con personal no técnico. Los percentiles constituyen una escala ordinal, permiten ordenar a las personas pero no garantizan la igualdad de intervalos, o, en otras palabras, diferencias iguales entre percentiles no implican diferencias iguales entre puntuaciones directas, constituyen una transformación no lineal de estas.

La realización concreta de la citada transformación para unos datos empíricos es muy sencilla, amén de que todos los programas de ordenador convencionales la llevan a cabo como parte de la descripción de los datos. Véase el ejemplo de la tabla 5.1,

TABLA 5.1

Puntuaciones en el test	Frecuencias absolutas	Frecuencias acumuladas (PM)	Percentiles (porcentaje)
10	2	199	99,50
9	4	196	98,00
8	12	188	94,00
7	10	177	88,50
6	46	149	74,50
5	50	101	50,50
4	40	56	28,00
3	16	28	14,00
2	10	15	7,50
1	6	7	3,50
0	4	2	1,00

en el que se pasó un test de 10 ítems a una muestra de 200 personas, obteniéndose la distribución de frecuencias que se detalla.

Para obtener los percentiles se obtienen primero las frecuencias acumuladas hasta el punto medio (columna 3) y luego se convierten en porcentajes mediante una sencilla regla de tres, con lo cual a cada puntuación directa (columna 1) le corresponde un percentil (columna 4).

No está justificado agrupar las puntuaciones directas en intervalos, como era costumbre en otras épocas con el fin de simplificar los entonces tediosos cálculos; si ello se hiciera, se perdería información, pues se asignaría la misma puntuación a todas las personas que caen en el mismo intervalo, cuando de hecho no la han obtenido necesariamente. Nótese también que la transformación de las puntuaciones directas en centiles no altera la forma de la distribución de las puntuaciones; compruébelo el lector representando gráficamente los datos de la tabla.

2. PUNTUACIONES TÍPICAS

Las puntuaciones directas pueden transformarse en otras denominadas «típicas» (Z_x) mediante una transformación lineal bien conocida, consistente en restarles la media y dividir por la desviación típica:

$$Z_x = \frac{X - \bar{X}}{S_x}$$

donde:

X: Puntuación directa.
$\bar{X}$: Media de la muestra.
S_x: Desviación típica.

Por tanto, la expresión anterior convierte automáticamente cualquier puntuación directa en típica. Aplicada a los datos de la tabla 5.1, donde la media es 4,91 y la desviación típica 1,866, se obtiene la escala típica de la tabla 5.2.

A la vista de la puntuación típica de una persona se sabe de inmediato si está por debajo (signo –) o por encima (signo +) de la media, así como el grado en que se separa de ella, pues la puntuación típica indica el número de desviaciones típicas que una persona se aparta de la media. Para un estudio detallado de las propiedades y comparabilidad de las escalas típicas, véase Amón (1984), limitándonos a señalar aquí que su media es cero y su desviación típica 1.

TABLA 5.2

Puntuaciones en el test (X)	Frecuencias absolutas	Puntuaciones típicas (Z_x)
10	2	2,7
9	4	2,2
8	12	1,7
7	10	1,1
6	46	0,6
5	50	0,1
4	40	–0,5
3	16	–1,0
2	10	–1,6
1	6	–2,1
0	4	–2,6

2.1. Típicas derivadas

El mayor inconveniente de tipo práctico para el uso de las típicas radica en los signos negativos y números decimales. Para evitarlo, las puntuaciones típicas se transforman a su vez en otras escalas que evitan estos dos inconvenientes, denominadas «típicas derivadas» (D).

Las típicas derivadas se obtienen a partir de las típicas primitivas mediante la transformación:

$$D = \bar{X}_D + S_D(Z_x)$$

donde:

$\bar{X}_D$: Media para la nueva escala derivada.
S_D: Desviación típica elegida para la nueva escala.
Z_x: Puntuación típica primitiva.

La media y la desviación típica elegidas son arbitrarias y solo obedecen a exigencias prácticas. Son muy populares, por ejemplo, las llamadas puntuaciones T de McCall, que ubican la media en 50 y la

desviación típica en 10, denominándose así, al parecer, en honor a Terman y Thorndike. No se prive el lector de instituir las suyas propias. Muchos test al uso utilizan este tipo de puntuaciones derivadas; por ejemplo, el WAIS ubica la media en 100 y la desviación típica en 15, el Standford-Binet en 100 y 16, respectivamente, y el MMPI en 50 y 10, siguiendo a McCall.

En la tabla 5.3 se han transformado las puntuaciones típicas anteriores en una escala derivada de media l00 y desviación típica 20.

TABLA 5.3

Puntuaciones en el test (X)	Frecuencias absolutas	Puntuaciones típicas (Z_x)	Típicas derivadas
10	2	2,7	154
9	4	2,2	144
8	12	1,7	134
7	10	1,1	122
6	46	0,6	112
5	50	0,1	102
4	40	–0,5	90
3	16	–1,0	80
2	10	–1,6	68
1	6	–2,1	58
0	4	–2,6	48

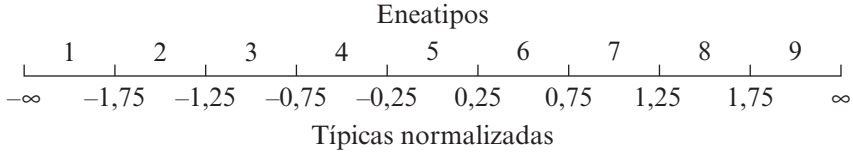
2.2. Típicas normalizadas

Las puntuaciones típicas normalizadas se obtienen a partir de los percentiles, buscando la puntuación típica correspondiente bajo la curva normal. Previamente a realizar esta transformación hay que comprobar estadísticamente, mediante χ^2 u otra prueba, que las puntuaciones empíricas se distribuyen según la curva normal; de lo contrario se estarían falseando los datos, al asumir una distribución que de hecho no se da empíricamente. Si la distribución se ajusta estrictamente a la normal, estas puntuaciones típicas normalizadas y las anteriores calculadas mediante la fórmula propuesta deberían resultar iguales. Dado que los datos empíricos raramente se distribuyen exactamente según la curva normal, esta transformación solo quedaría legitimada si el ajuste resulta estadísticamente aceptable. Véanse en la tabla 5.4 las típicas normalizadas correspondientes al ejemplo que venimos utilizando.

TABLA 5.4

Puntuaciones en el test (X)	Frecuencias absolutas	Puntuaciones típicas (Z_x)	Típicas normalizadas
10	2	2,7	2,57
9	4	2,2	2,05
8	12	1,7	1,56
7	10	1,1	1,20
6	46	0,6	0,66
5	50	0,1	0,01
4	40	–0,5	–0,58
3	16	–1,0	–1,08
2	10	–1,6	–1,44
1	6	–2,1	–1,81
0	4	–2,6	–2,33

A partir de las típicas normalizadas se pueden obtener todo tipo de puntuaciones derivadas, análogamente a como ocurría con las típicas primitivas, denominándose ahora, por razones obvias, «derivadas normalizadas». Una de estas escalas derivadas son los afamados estaninos o eneatipos. Su nombre procede de *standard nine,* pues se divide el rango total de las puntuaciones en nueve intervalos y se asigna a cada uno un número del 1 al 9. Los límites de tales intervalos se fijan según el gráfico:



Es una escala poco recomendable, pues se pierde mucha información al incluir en el mismo eneatipo personas con distinta puntuación. Su única ventaja es la facilidad para ser entendida por los no expertos. Nótese que en realidad es una escala derivada de media 5 y desviación típica 2:

$$\text{Eneatipos} = 5 + 2(Z_n)$$

siempre y cuando se tenga en cuenta que el valor máximo es 9, luego a todo valor superior se le asigna el 9, y a todo valor inferior a 1 se le asigna el 1, redondeando los valores intermedios para obtener el eneatipo correspondiente.

Por ejemplo, averiguar el eneatipo correspondiente a las siguientes típicas normalizadas:

$$Z_1 = 0{,}7, \quad Z_2 = 2{,}5 \quad \text{y} \quad Z_3 = -3{,}5$$

Una ojeada al gráfico pone de manifiesto que los eneatipos buscados son, respectivamente, 6, 9 y 1. Aplicando la fórmula:

$$E_1 = 5 + 2(0{,}7) = 6{,}4 \qquad \text{(Eneatipo 6)}$$
$$E_2 = 5 + 2(2{,}5) = 10 \qquad \text{(Eneatipo 9)}$$
$$E_3 = 5 + 2(-3{,}5) = -2 \qquad \text{(Eneatipo 1)}$$

3. EDAD

Otra posible transformación de las puntuaciones directas es en edades. En líneas generales, para llevar a cabo la transformación se administra previamente el test a personas de diferentes edades, calculando la media del test para cada edad. La regla de transformación será asignar a cada persona la edad que le corresponda por su puntuación. Por ejemplo, si los niños de 7 años obtienen en el test una media de 20 puntos, cuando un niño, independientemente de su edad cronológica, obtenga 20 puntos en ese test se le asignará una edad de 7 años. Obviamente las escalas suelen ser más detalladas e ir de seis meses en seis meses, o incluso menos, de mes en mes, por ejemplo.

Este tipo de escalamiento por edades se presta a numerosas confusiones e interpretaciones erróneas, y se utiliza cada vez menos. Tendría más sentido para escalar atributos que crecen o decrecen sistemáticamente con la edad, pero en la mayoría de las variables de interés psicológico no es el caso.

Nótese que originalmente el cociente intelectual (CI) se basaba en este tipo de escala:

$$CI = \frac{EM}{EC} \times 100$$

al dividir la edad asignada a la persona según su puntuación en el test (edad mental) entre la edad cronológica, multiplicando por 100 para evitar los decimales.

El inconveniente principal de este tipo de escalas de edades radica en que si la función psicológica medida no aumenta o disminuye linealmente con la edad, como suele ocurrir, las diferencias de edad mental asignadas no tienen el mismo significado para todas las edades cronológicas. A ello se añade además el llamado «efecto de techo», consistente en que a partir de cierta edad (techo) la función psicológica medida no aumenta significativamente, y por tanto carece de sentido utilizar la edad como unidad de medida. Por si fuera poco, se produce, además, otro efecto colateral de interpretación poco deseable, tendiendo a equipararse en términos generales a los sujetos a los que se asigna la misma edad en determinada variable psicológica, cuando el funcionamiento cognitivo global puede ser muy diferente.

Se han citado aquí algunos de los tipos de transformaciones más habituales, pero otras muchas, seguramente infinitas, son posibles; el uso de una u otra dependerá de las exigencias prácticas de cada situación.

Señalar, finalmente, que no es infrecuente denominar «baremo» al conjunto formado por las puntuaciones directas en el test y las correspondientes transformadas por alguno de los métodos descritos, u otros. Asimismo, suele denominarse «grupo normativo» a la muestra de personas utilizada para establecer las transformaciones que se pretende sean válidas para todas las personas de la población de donde se extrajo la muestra. Nótese que si por alguna razón la muestra está incorrectamente elegida, por ejemplo, no es aleatoria o contiene un número insuficiente de personas, los juicios que se hagan basándose en las normas (transformaciones) elaboradas a partir de ella serán incorrectos. Un buen ejemplo de estas deficiencias puede observarse echando una ojeada a los baremos que aparecen en los manuales de algunos de los test editados en España, en los que se utilizan muestras con un número de personas insuficiente, o no se actualizan periódicamente, habiéndose quedado obsoletos y no representando la situación actual de la población.

Un análisis detallado y clásico sobre la transformación de las puntuaciones y la casuística posible en la elección y descripción de grupos normativos puede consultarse en Angoff (1984).

EJERCICIOS

1. En la tabla adjunta aparecen las puntuaciones obtenidas por cinco personas en un test de cuatro ítems.

Personas	Ítems			
	1	2	3	4
A	1	0	0	0
B	1	1	0	0
C	1	1	1	0
D	1	1	1	1
E	0	0	0	0

Expresar la puntuación directa del sujeto C en:

1. Puntuaciones típicas.
2. Puntuaciones típicas derivadas de media 100 y desviación típica 10.
3. Percentiles.
4. Típicas normalizadas.
5. Eneatipos.
6. CI de desviación (media 100 y desviación típica 15).

2. Aplicado un test de inteligencia general a una muestra de 1.000 sujetos, las puntuaciones se distribuyeron según la curva normal con media 20 y desviación típica 5.

1. Uno de los sujetos obtuvo en el test una puntuación directa de 15 puntos; exprese dicha puntuación en escala:

   — Diferencial.
   — Típica.
   — Eneatipos.
   — Percentiles.
   — Derivada de media 88 y desviación típica 6.

2. Otro de los sujetos «solo» fue superado por 900 de sus compañeros; expresar su puntuación en todas las escalas anteriores, incluida la de puntuaciones directas.

3. Jensen y Munro (1979) encontraron que para una muestra de 39 mujeres el tiempo de reacción medio a estímulos visuales era de 330 milisegundos. Asumiendo que los tiempos de reacción se distribuyen según la curva normal con una desviación típica de 20.

1. ¿Qué puntuación directa, diferencial, típica, eneatipo, percentil y derivada con media 100 y varianza 225 le correspondería a una de las mujeres cuyo tiempo de reacción fue de 370 milisegundos?

4. A una oferta de empleo se presentaron 1.000 aspirantes a los que se les aplicó un test de selección. Las puntuaciones en este test se distribuyeron según la curva normal con una media de 40 y una desviación típica de 10. Solo fueron admitidos para continuar el proceso de selección aquellos 100 que obtuvieron las mejores puntuaciones en el test.

En relación con el sujeto que obtuvo la puntuación más baja de entre los admitidos, calcular:

1. Su percentil.
2. Su puntuación típica.
3. Su puntuación directa.
4. Su puntuación diferencial.
5. Su eneatipo.
6. Su puntuación derivada en una escala de media 60 y desviación típica 20.
7. Si del enunciado del problema se suprimiese «las puntuaciones en este test se distribuyeron según la curva normal», ¿cuáles de los apartados del problema podría usted contestar? Razone la respuesta.

5. En la tabla adjunta aparecen las puntuaciones obtenidas por 400 sujetos en un test de 50 ítems.

Puntuaciones en el test	Frecuencias absolutas
50	4
45	8
40	24
35	20

Puntuaciones en el test	Frecuencias absolutas
30	92
25	100
20	80
15	32
10	20
5	12
0	8

Elabore los baremos del test en:

1. Percentiles.
2. Puntuaciones típicas.
3. Puntuaciones derivadas de media 100 y desviación típica 20.
4. Típicas normalizadas.

SOLUCIONES

1.1. 0,71
2. 107
3. 70
4. 0,52
5. 6
6. 108

2.1. –5, –1, 3, 16, 82
2. 13,6, –6,4, –1,28, 2, 10, 80

3.1. 370, 40, 2, 9, 98, 130

4.1. 90
2. 1,28
3. 52,8
4. 12,8
5. 8
6. 86
7. 4.1

5.

Percentiles	Típicas	Derivadas	Normalizadas
99,50	2,7	154	2,57
98,00	2,2	144	2,05
94,00	1,7	134	1,56
88,50	1,1	122	1,20
74,50	0,6	112	0,66
50,50	0,1	102	0,01
28,00	–0,5	90	–0,58
14,00	–1,0	80	–1,08
7,50	–1,6	68	–1,44
3,50	–2,1	58	–1,81
1,00	–2,6	48	–2,33

6 Equiparación de las puntuaciones

La equivalencia o equiparación de puntuaciones *(equating)* de dos o más test se refiere al establecimiento de una correspondencia entre las puntuaciones de uno y otro, de tal modo que sea indiferente cuál se aplique a las personas, pues sus puntuaciones en uno serán expresables en términos del otro, si efectivamente la mentada equiparación se ha hecho adecuadamente. En palabras de Angoff (1982a), la equiparación de las puntuaciones es el proceso de desarrollar una conversión del sistema de unidades de un test al sistema de unidades de otro, de tal modo que las puntuaciones derivadas de ambos test después de la conversión sean equivalentes o intercambiables.

El problema de la equiparación de las puntuaciones nunca fue un tema al que la psicometría clásica prestara gran atención, salvo, como señala Brennan (1987), aquellos psicómetras con responsabilidades directas en las grandes compañías constructoras de test. Efectivamente, Gulliksen (1950) lo trata de pasada, Lord y Novick (1968) apenas lo citan y los «Standards for Educational and Psychological Testing» de 1974 ni lo mentan. El tratamiento pionero y clásico es el de Angoff incluido en el libro editado por Thorndike (1971), capítulo que en 1984 editará en forma de libro el Educational Testing Service (ETS). En 1982, también bajo los auspicios del ETS, se publica un libro monográfico sobre el tema (Holland y Rubin, 1982) al que contribuyen numerosos especialistas y en el que se incluye una bibliografía exhaustiva de lo hecho hasta entonces. Por los años ochenta los trabajos son abundantes: Lord (1980) le dedica un capítulo, y los «Standards for Educational and Psychological Testing» (1985), varios «standards». En 1987 la revista *Applied Psychological Measurement* edita un número especial al respecto, y, por su parte, «Educational Measurement» incluye en su sección instructiva «Ítems» una exposición muy asequible y divulgativa de Kolen (1988) sobre metodología clásica para la equiparación. Tratamientos más recientes y detallados pueden verse en Von Davier (2011) o Kolen y Brennan (2014), y en español, en Navas (1996).

Las causas de esta eclosión hay que buscarlas, en primer lugar, en el uso masivo de los test en Estados Unidos, con repercusiones tan relevantes como quién puede acceder (y dónde) a la enseñanza universitaria, empleos, promociones, certificaciones, etc. Ello obliga a los constructores a elaborar varias formas del mismo test sucesivamente, con el consiguiente problema implicado de comparar y equiparar las puntuaciones obtenidas en ellos, so pena de graves injusticias comparativas. La continua crítica y discusión social de este sistema generalizado de test ha obligado a los constructores a justificar y explicar públicamente sus métodos de equiparación. En segundo lugar, los nuevos modelos de TRI que dominan la psicometría actual permiten un tratamiento más adecuado del problema de la teoría clásica, véase el epígrafe 9 del capítulo 7.

Para hablar propiamente de establecer una equiparación entre las puntuaciones de dos test, ambos han de medir la misma variable y con la misma fiabilidad. Si se trata de variables distintas, el concepto de equivalencia carece de sentido, aunque nada impediría intentar predecir una a partir de la otra, eso es otra cuestión. Respecto a la misma fiabilidad, de no darse se seguiría la inaceptable posibilidad de equiparar un test con otro menos fiable y asumir que las puntuaciones en ambos son inter-

cambiables. Una exposición detallada de las condiciones teóricas exigibles para establecer equiparaciones rigurosas puede consultarse en Lord (1980), quien demuestra, ironías del destino, que una equiparación estricta solo es posible cuando es innecesaria. No obstante, en el trato con la sucia realidad los métodos que se comentan brevemente a continuación no son del todo desatinados.

Suele hablarse de equiparación horizontal cuando los test a equiparar se intentan construir a priori con igual dificultad, caso, por ejemplo, de las formas alternativas del mismo test. Si la dificultad de los test a equiparar es claramente distinta, se habla de equiparación vertical. Caso típico de lo cual es cuando se desea establecer comparaciones entre competencias que aumentan con la edad, utilizando test de diferente nivel (dificultad) a cada edad. En realidad es un problema típico de escalamiento y se presenta en la práctica asociado con competencias escolares en relación con el curso o edad de los estudiantes.

1. DISEÑOS

Tres han sido los diseños más frecuentemente utilizados desde la óptica clásica:

— Un solo grupo.
— Dos grupos equivalentes.
— Test de anclaje.

Un solo grupo

Supóngase que se dispone de dos test cuyas puntuaciones se desea equiparar. En el diseño de un solo grupo se elegiría una muestra aleatoria de personas a las que se aplicarían ambos test y posteriormente, por alguno de los métodos que luego se citan, se procedería a la equiparación. Para mitigar los efectos del orden de aplicación de los test suelen aplicarse contrabalanceados, es decir, se divide la muestra en dos partes aleatorias y en cada una de ellas se aplican los dos test a equiparar en distinto orden.

Dos grupos equivalentes

En este diseño se eligen dos muestras aleatorias de la población a la que se destinan los dos test y se aplica uno de los test a cada una de ellas, procediéndose luego a la equiparación, bajo el supuesto de que el azar generó muestras de sujetos equivalentes.

Test de anclaje

Ha sido (y es) el diseño más utilizado. Se aplican los dos test a equiparar a dos muestras (uno a cada una), como en el caso anterior, pero además a ambas se le aplica cierto número de ítems comunes de anclaje que permitirán establecer las equivalencias entre los test a equiparar. Nótese que aquí las dos muestras no tienen por qué ser necesariamente equivalentes.

2. MÉTODOS

Los tres métodos más utilizados por la psicometría clásica para establecer las equiparaciones son:

— Media.
— Transformación lineal.
— Percentiles.

Media

Es el método más elemental de equiparar, que consiste en hacer corresponder las medias de los test a equiparar. Por ejemplo, si la media de un test X es 40 y la de otro Y es 45, la equiparación se establecería sumando 5 puntos a las puntuaciones de los sujetos en el test X, o, lo que es lo mismo, restando 5 a las del test Y. Es un método simplísimo pero con fuertes asunciones acerca de la distribuciones de las puntuaciones para que la equiparación tenga algún sentido, pero prácticamente no se usa en la actualidad.

Transformación lineal

Consiste en equiparar las puntuaciones típicas. Se equiparan aquellas puntuaciones directas con típicas iguales. Para dos test X e Y:

$$Z_x = Z_y$$

explícitamente:

$$\frac{X - \bar{X}}{S_x} = \frac{Y - \bar{Y}}{S_y}$$

despejando Y:

$$Y = \frac{S_y}{S_x}(X - \bar{X}) + \bar{Y} \qquad [6.1]$$

Es decir, las puntuaciones del test Y equivalentes a las del test X vienen dadas por una transformación lineal de estas. Por ejemplo, si una muestra aleatoria de sujetos obtiene en el *test X* de compresión verbal una media de 40 y una desviación típica de 4 y en otro test Y, también de comprensión verbal, una segunda muestra aleatoria (diseño de grupos equivalentes), obtiene una media de 60 y una desviación típica de 6, ¿qué puntuación en el test Y sería equivalente a una puntuación de 48 en el test X?

$$\frac{48 - 40}{4} = \frac{Y - 60}{6}$$

despejando Y:

$$Y = \frac{6}{4}(48 - 40) + 60 = 72$$

Obtener 48 puntos en X es equiparable a obtener 72 en Y. La aplicación de este método de transformación lineal al diseño de grupos equivalentes se hace tal como se acaba de exponer; su implementación con algunos matices para los otros dos diseños se comenta a continuación.

Según Angoff (1984), el error típico de medida para las puntuaciones así equiparadas viene dado por:

$$S_e = \sqrt{\frac{2S_y^2}{N_t}(Z_x^2 + 2)} \qquad [6.2]$$

donde N_t es el número total de sujetos (sumados los de ambas muestras) y $Z_x = (X - \bar{X})/S_x$.

Nótese cómo este error típico aumenta con el valor de Z_x, esto es, a medida que las puntuaciones equiparadas X se alejan de la media, el error típico de la equiparación es mayor.

En el diseño de un solo grupo, como ya se ha señalado, los test suelen aplicarse en distinto orden a cada una de las submuestras en las que se divide la muestra, para contrabalancear los posibles efectos del orden de aplicación. Ello complejiza ligeramente el cálculo de los datos necesarios en el método de transformación lineal. En concreto, los valores globales de la muestra: $\bar{X}$, $\bar{Y}$, S_x, S_y han de obtenerse a partir de los valores de las submuestras, mediante las técnicas estadísticas al uso (véase, por ejemplo, Amón, 1984), dado que cada test se aplicó a cada submuestra, en distinto orden. Una vez hecho esto, se procede análogamente al caso anterior de dos grupos.

El error típico de medida de las puntuaciones equiparadas (Angoff, 1984) aquí viene dado por:

$$S_e = \sqrt{\frac{S_y^2(1 - r_{xy})[Z_x^2(1 + r_{xy}) + 2]}{N_t}} \qquad [6.3]$$

Nótese que este error típico es menor que el dado para el diseño de dos grupos equivalentes; para conseguir la misma precisión con ambos aquel requiere más sujetos que este.

En el caso del diseño con test de anclaje, como se ha visto, se dispone de dos muestras, A y B, y se aplica a cada una de ellas uno de los test a equiparar, sea el test X a A y el test Y a B. Además, un test Z se aplica en ambas muestras. Los datos necesarios para el método de transformación lineal (Angoff, 1984) pueden obtenerse del siguiente modo:

$$\bar{X} = \bar{X}_A + b_{XZ(A)}(\bar{Z} - \bar{Z}_A)$$
$$\bar{Y} = \bar{Y}_B + b_{YZ(B)}(\bar{Z} - \bar{Z}_B)$$
$$S_x = \sqrt{S_{XA}^2 + b_{XZ(A)}^2(S_Z^2 - S_{ZA}^2)}$$
$$S_y = \sqrt{S_{YB}^2 + b_{YZ(B)}^2(S_Z^2 - S_{ZB}^2)}$$

donde

$b_{XZ(A)}$: Pendiente de X sobre Z en la muestra A.
$b_{YZ(B)}$: Pendiente de Y sobre Z en la muestra B.

$\bar{Z}$: Media global del test de anclaje.
S^2_{XA} y S^2_{YB}: Varianzas respectivas de X e Y.

Obtenidos de ese modo los valores $\bar{X}$, $\bar{Y}$, S_x, S_y, se procede análogamente a los casos anteriores.

Para este diseño el error típico de medida viene dado por:

$$S_e = \sqrt{\frac{2S_y^2(1 - r^2)[(1 + r^2)Z_x^2 + 2]}{N_t}} \qquad [6.4]$$

donde se asume que:

$$r = \frac{b_{XZ(A)}}{S_x} = \frac{b_{YZ(B)}}{S_y}$$

Percentiles

Es el método más habitual; tanto es así que a veces se han definido como puntuaciones equivalentes aquellas con percentiles iguales. El método consiste en eso, en hacer corresponder o equiparar aquellas puntuaciones de ambos test cuyos percentiles son iguales. Por ejemplo, si en un test X de CV a una puntuación directa de 23 le corresponde el percentil 80 y en otro test Y, también de CV, a una puntuación de 25 le corresponde asimismo el percentil 80, la puntuación de 23 en el test X se equipara a 25 en el test Y; sacar 23 puntos en X equivale según este método a sacar 25 en Y. Las posibilidades y limitaciones de este método son aquellas inherentes a los percentiles.

7 Teoría de respuesta a los ítems

La teoría de respuesta a los ítems (TRI) constituye un nuevo enfoque en la teoría de los test que permite resolver ciertos problemas de medición psicológica inatacables desde la teoría clásica de los test (TCT). Como señala Lord (1980), la TRI no contradice ni las asunciones ni las conclusiones fundamentales de la teoría clásica de los test, sino que hace asunciones adicionales que permitirán responder cuestiones que la TCT no podía. No obstante, como se irá viendo, la TRI constituye un giro importante en el acercamiento a la medición psicológica, y, como el propio Lord (1980) indica a continuación, a pesar de este carácter complementario de la TRI respecto de la TCT, poco de esta se utilizará explícitamente en su formulación.

El nombre «teoría de respuesta a los ítems» proviene de que este enfoque se basa en las propiedades de los ítems más que en las del test global. Aunque ha sido frecuente en el pasado referirse a la TRI como teoría o modelos de rasgo latente, en la actualidad la denominación universal es TRI. Y ello, efectivamente, porque refleja el funcionamiento real de estos modelos basados en los ítems, permitiendo además distinguirlos de otros acercamientos más generales que utilizan el concepto de rasgo latente en psicología, como pueden ser el análisis factorial, el análisis multidimensional o las ecuaciones estructurales (Hambleton y Swaminathan, 1985).

Los orígenes de la TRI hay que buscarlos en los trabajos pioneros de Richardson (1936), Lawley (1943), Tucker (1946), Lord (1952, 1953a, 1953b) y Birnbaum (1957, 1958a, 1958b), produciéndose una rápida expansión a partir de los años sesenta con la aparición del libro de Rasch (1960) y, sobre todo, con las contribuciones de Birnbaum en el de Lord y Novick (1968), todo ello complementado con el acceso generalizado a los ordenadores, imprescindibles para el tratamiento de los modelos de TRI. En la actualidad se dispone de una extensa y pertinente literatura, por ejemplo Wright y Stone (1979), Lord (1980), Hambleton y Swaminathan (1985), Andrich (1988), Hulin, Drasgow y Parsons (1983), Van der Linden y Hambleton (1997), Embretson y Reise (2000), Yen y Fitzpatrick (2006), Ayala (2009), Nering y Ostini (2010), Faulkner-Bond y Wells (2016) o el reciente tratamiento enciclopédico de Van der Linden (2016), entre otros muchos. En español pueden verse López Pina (1995), Muñiz (1996a, 1997a), Revuelta, Abad y Ponsoda (2006), Martínez Arias et al. (2006) o Abad et al. (2011).

1. OBJETIVOS

Aparte de las contribuciones de tipo técnico que aportará la TRI a la hora de construir y analizar los test, desde el punto de vista teórico de la medición su gran contribución se centra en la posibilidad de obtener mediciones invariantes respecto de los instrumentos utilizados y de las personas implicadas. Veamos qué significa esta afirmación un tanto críptica. En la teoría clásica, el resultado de la medición de una variable depende del test utilizado, lo que plantea serios problemas para tratar de establecer la equivalencia entre las puntuaciones de dos test distintos que midan una misma variable. Por ejemplo, si la variable inteligencia de una persona se mide con dos test distintos, ambos de inteligencia obviamente, su puntuación no será la misma en cada uno de ellos, ya que no necesariamente funcio-

nan en la misma escala. Por tanto ¿cuál es la inteligencia de esa persona? En la teoría clásica la medida de una variable es inseparable del instrumento utilizado para medirla, y ello constituye una seria limitación. El problema no es ni mucho menos nuevo, pues Thurstone (1928b) ya lo apuntó claramente: «... un instrumento de medida no debe venir afectado por los objetos medidos... sus mediciones deben ser independientes de los objetos medidos» (p. 547). Además, las propiedades del instrumento de medida, esto es, de los ítems y, por tanto, del test, están en función de las personas a las que se aplican. Por ejemplo, el índice de dificultad de un ítem dependerá de que el grupo de personas utilizado para calcularlo sea competente o no; en el primer caso será fácil, y en el segundo, difícil. Para entender estas dos limitaciones, imagínese el lector que la longitud de una mesa dependiese del tipo de metro utilizado y que, además, las cualidades del metro se estableciesen en función del tipo de mesa medida. El acercamiento clásico se encontraba encerrado en esa incongruencia teórica: la medición depende del instrumento utilizado y las propiedades de estos están en función de los objetos medidos, las personas. La promesa y objetivo central de la TRI serán solucionar este problema, lo que, en suma, permitirá:

— Obtener mediciones que no varíen en función del instrumento utilizado, es decir, que sean invariantes respecto de los test empleados.
— Disponer de instrumentos de medida cuyas propiedades no dependan de los objetos medidos, sean invariantes respecto de las personas evaluadas.

Además de este objetivo central, o, más bien, derivados de él, la TRI proporcionará todo un conjunto de avances técnicos de gran interés para la evaluación psicológica y educativa, tales como las funciones de información de los ítems y del test, errores típicos de medida distintos para cada nivel de la variable medida o el establecimiento de bancos de ítems con parámetros estrictamente definidos, lo que posibilita el uso de test adaptados al nivel de la persona evaluada, permitiendo exploraciones más exhaustivas y rigurosas en función de las características de las personas. Para conseguir tales objetivos las asunciones de la TRI serán fuertes y restrictivas, amén de comprometidas desde el punto de vista de la teorización psicológica.

2. SUPUESTOS

2.1. Curva característica de los ítems

Los modelos de TRI asumen que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar estos y denominan a dicha función «curva característica del ítem» (CCI). Expresado en otras palabras, ello significa que la probabilidad de acertar un ítem depende de los valores de la variable medida por el ítem; por tanto, personas con distinta puntuación en dicha variable tendrán probabilidades distintas de superar determinado ítem.

En la figura 7.1 aparece la curva característica de un ítem. En el eje de abscisas se representan los valores de la variable que mide el ítem, a la que en adelante se denominará theta (θ), y en el de ordenadas aparece la probabilidad de acertar el ítem $P(\theta)$. La curva dibujada es la CCI y nos da la probabilidad de acertar el ítem para los distintos valores de θ. Así, por ejemplo, las personas cuyo valor en θ fuese –1 tendrían una probabilidad de acertar este ítem de 0,10, a una $\theta = 0$ le correspondería una $P(\theta) = 0{,}50$, etc.

La CCI, como su nombre indica, es eso, característica, típica, específica de cada ítem, caracteriza al ítem; por tanto, las CCI de los ítems que miden una determinada variable θ no son iguales, si bien compartirán determinada forma general, como se verá más adelante. Llegados aquí seguramente asaltarán al lector varios interrogantes; el primero, cómo se elaboran las CCI y qué formas toman, pues posibles curvas en el plano hay infinitas, sin entrar a considerar espacios de más dimensiones; el segundo, cómo se relacionan, qué tienen que ver las CCI con los objetivos de la TRI establecidos. Por el momento solo cabe dar ánimos para continuar, todo llegará.

En primer lugar, una aclaración: la CCI no es la regresión ítem-test, aunque tenga algunas semejanzas. La regresión ítem-test consiste en hacer corresponder los valores del test con las proporciones de aciertos en determinado ítem.

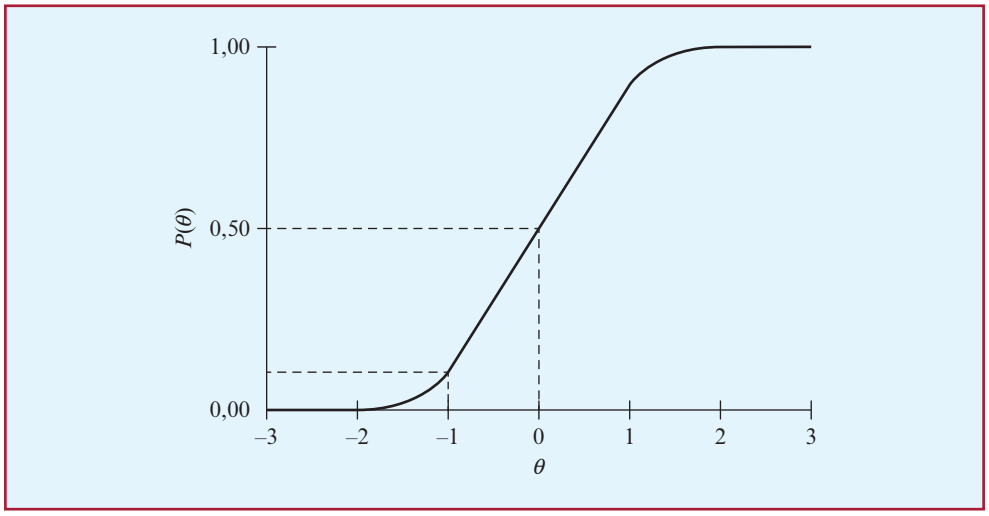


Figura 7.1.—Curva característica de un ítem.

En la figura 7.2 aparece la regresión de un ítem sobre el test para una determinada muestra de personas. En el eje de abscisas se representan las puntuaciones de las personas en el test, y en el de ordenadas, la proporción de personas que acertaron elítem. Los puntos del gráfico indican la proporción de personas que aciertan el ítem para cada valor del test. Por ejemplo, de las personas que sacaron 8 puntos en el test acertaron el ítem el 70 % (p = 0,70), mientras que de las que sacaron un 1 en el test solo lo acertaron el 20 % (p = 0,20).

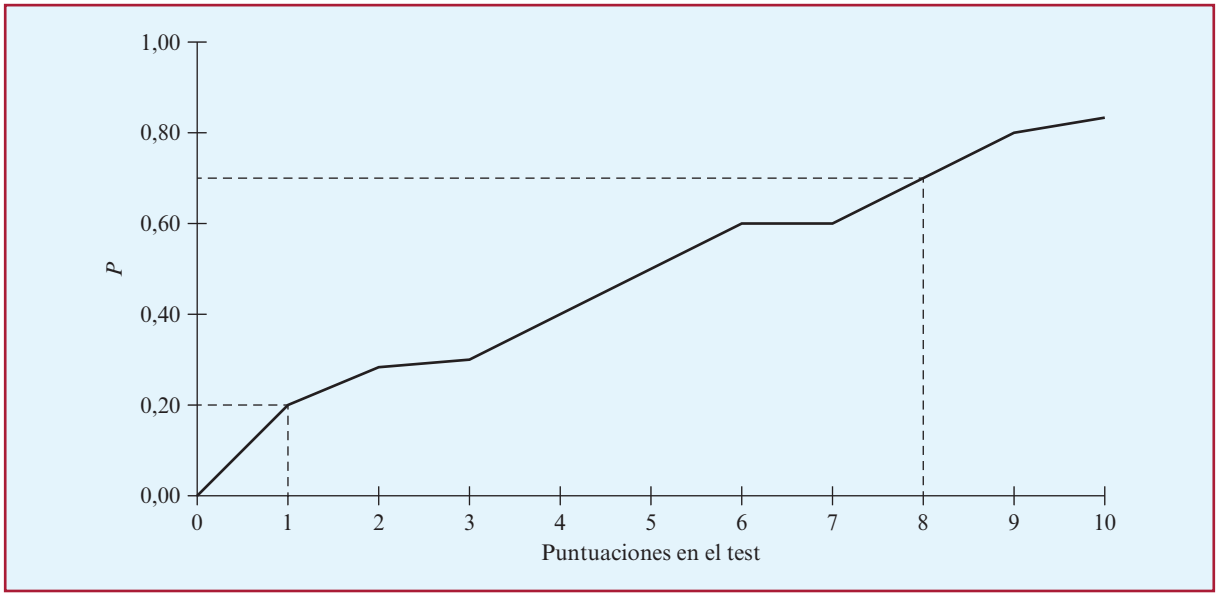


Figura 7.2.—Regresión ítem-test.

La diferencia fundamental entre la CCI y la regresión ítem-test es que en la CCI la variable que miden los ítems (θ) no es la puntuación que las personas sacan en el test, lo cual no quiere decir que no estén relacionadas, pero no son intercambiables sin más. Por ejemplo, los valores de θ están comprendidos entre $-\infty$ y $+\infty$, mientras que los de un test suelen estarlo entre cero y la puntuación máxima posible en ese test. Puede decirse que las puntuaciones de las personas en el test son una estimación de θ, pero no constituyen la escala θ. La forma estricta de conexión entre las puntuaciones del test y los valores de θ se verá más adelante.

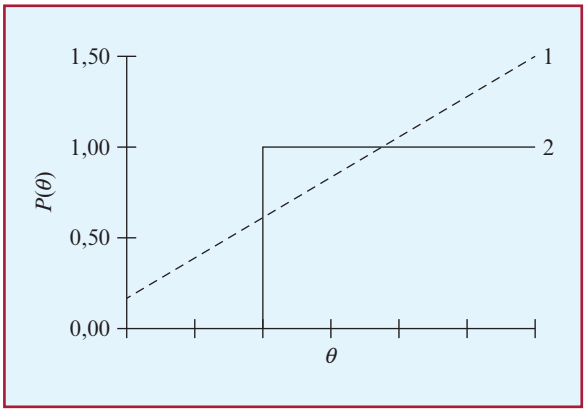


Figura 7.3.—Tipos de curvas características no utilizados en la teoría de respuesta a los ítems.

Parámetros

Los parámetros se refieren a los valores que van a hacer que la CCI tenga una forma u otra. Habrá que especificar en primer lugar cuántos se toman en consideración y en segundo lugar qué valores toman, es decir, cómo se calcula su valor.

Veamos en principio, a modo de ilustración, dos tipos de CCI que no se van a utilizar en la TRI.

Si se utilizase como CCI la recta, ocurriría que, para determinados valores de θ, la probabilidad $P(\theta)$ podría ser negativa o mayor que 1, lo que sería incompatible con los axiomas de la probabilidad, que establecen su valor entre 0 y 1. En el caso 2 no se plantea el problema anterior, pero el modelo es poco plausible para modelizar la conducta humana, en la que el paso fallar-acertar un ítem no se produce tan sistemáticamente en un punto concreto y siempre el mismo, más bien se da una transición probabilística. Recuerde el lector que las clásicas escalas de Guttman se basan en este tipo de modelo

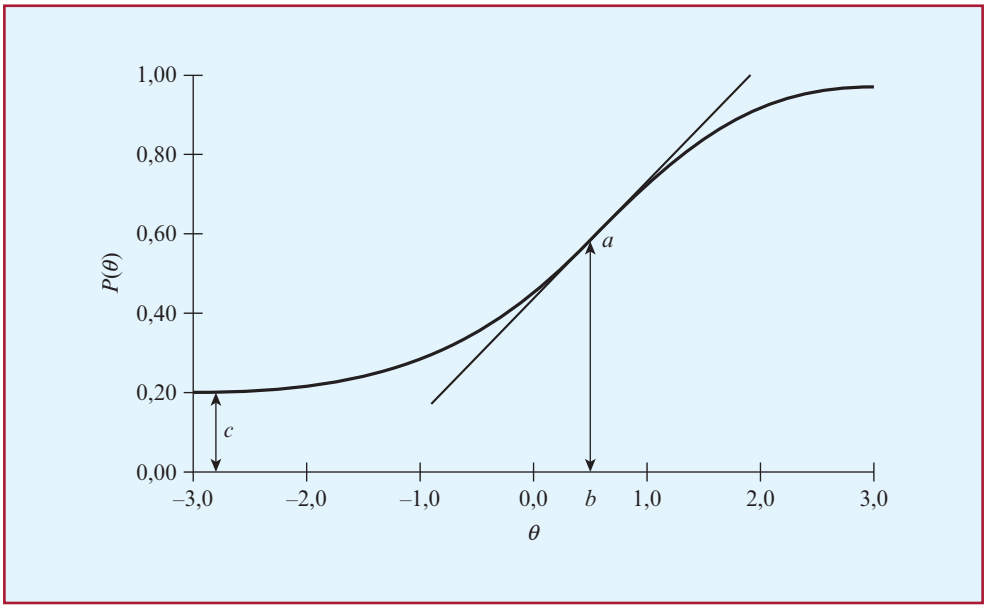


Figura 7.4.—Parámetros de las curvas características de los ítems.

determinista de todo o nada, poco plausible para humanos mayormente instalados en el reino de la probabilidad.

Las CCI utilizadas en la TRI van a ser del «tipo S» y para definirlas adecuadamente habrá que tener en cuenta únicamente los tres parámetros que se detallan a continuación.

Los tres parámetros a considerar van a denominarse *a*, *b* y *c*.

Parámetro a

El parámetro *a* se denomina *índice de discriminación* y su valor es proporcional a la pendiente de la recta tangente a la CCI en el punto de máxima pendiente de esta. Cuanto mayor sea la pendiente, mayor será el índice de discriminación. Su valor numérico se especificará más adelante, cuando se adopte un tipo de función matemática para la curva. Aunque el nombre «índice de discriminación» alude como en la teoría clásica a la capacidad discriminativa del ítem, su valor no será el mismo. A modo de ilustración, y para ayudar a su comprensión, cabe señalar (Lord, 1980) que el valor de *a* cuando θ se distribuye según la curva normal con media 0 y desviación típica 1 $N(0,1)$, y no hay aciertos al azar ($c = 0$), viene dado aproximadamente por la expresión:

$$a \equiv \frac{r_b}{\sqrt{1-(r_b)^2}} \qquad [7.1]$$

donde r_b es la correlación biserial ítem-test, o sea, el índice de discriminación en el modelo clásico.

Parámetro b

Se denomina *índice de dificultad* y es el valor de θ correspondiente al punto de máxima pendiente de la CCI. Como en el caso de *a,* tampoco el significado de *b* es aquí exactamente el mismo que en la teoría clásica, aunque, por supuesto, se refiere a la dificultad del ítem. Nótese, sobre todo, que aquí la dificultad del ítem se mide en la misma escala que θ; de hecho, es un valor de θ, aquel que corresponde a la máxima pendiente de la CCI. De nuevo, con fines ilustrativos, si se mantienen las condiciones de normalidad de θ, el parámetro *b* se relaciona con el índice de dificultad de la teoría clásica aproximadamente según la expresión:

$$b \equiv \frac{-Z_p}{r_b} \qquad [7.2]$$

donde Z_p es la puntuación típica que corresponde en la curva normal a la proporción de aciertos en el ítem (índice de dificultad en la teoría clásica) y r_b es la correlación biserial ítem-test. Por ejemplo, si en las condiciones citadas cierto ítem tiene una $r_b = 0{,}50$ y es acertado por el 75% de las personas (índice de dificultad en la teoría clásica, 0,75), *b* valdría aproximadamente:

$$b \equiv -0{,}67/0{,}50 = -1{,}34$$

(Nótese que el valor de *Z* en la curva normal que deja por encima de sí al 75% de los casos es –0,67.)

Parámetro c

El parámetro *c* representa la probabilidad de acertar el ítem al azar cuando «no se sabe nada», es decir, es el valor de $P(\theta)$ cuando $\theta = -\infty$. En otras palabras, es el valor asintótico de la CCI cuando θ tiende a $-\infty$. Su equivalente aproximado en la teoría clásica viene dado por la probabilidad de acertar el ítem al azar. Por ejemplo, si un ítem tiene cinco alternativas y solo una es correcta, la probabilidad de acertarlo al azar sin conocer la respuesta es $p = 1/5 = 0{,}20$. Luego para este ítem una estimación aproximada de *c* sería 0,20.

La CCI queda definida cuando se especifica el valor de estos tres parámetros y se adopta una determinada función matemática para la curva. Según el tipo de función matemática adoptada y el valor de los parámetros, tendremos diferentes modelos de CCI. Véanse en la figura 7.5 algunas CCI en las que los tres parámetros expuestos toman diferentes valores.

A medida que las CCI se ubican más a la derecha en el eje de abscisas, significa que los ítems son más difíciles, pues *b* aumenta. El ítem más fácil de los cinco representados es el 1, y el más difícil, el 5. El poder discriminativo *a* viene indicado por la pendiente de las CCI; los ítems 1 y 2 tienen un alto

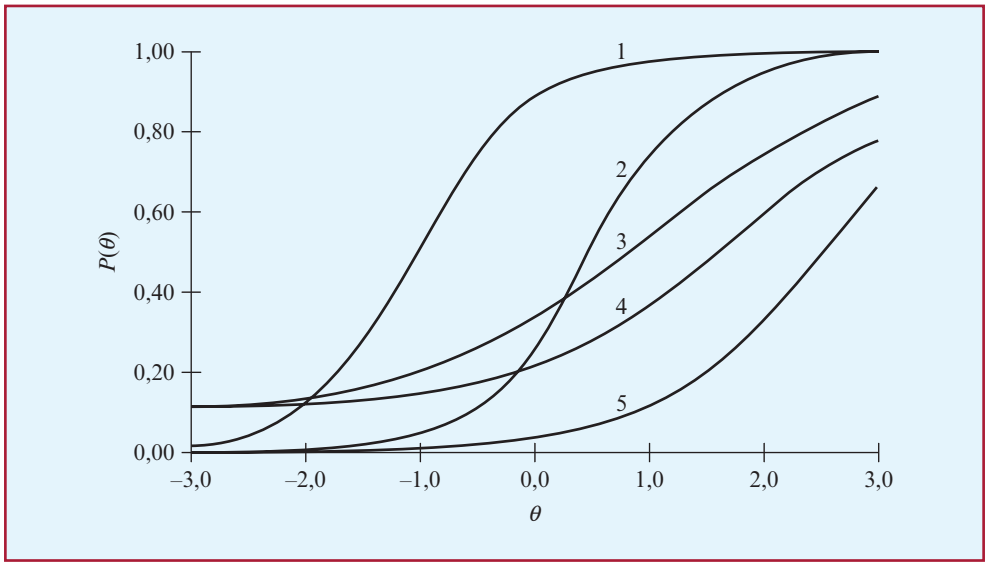


Figura 7.5.—Curvas características con diferentes valores de los parámetros.

poder discriminativo, pero hay que notar inmediatamente que esa capacidad discriminativa se da para determinados valores de θ, en concreto, en el ítem 1, para valores de θ en torno a –1, y en el ítem 2, para valores de θ en torno a 0,50. Esto tendrá importantes consecuencias en la construcción de test, pues, según nos interese discriminar en una zona u otra de θ, elegiremos unos ítems u otros. Finalmente, el parámetro c, aciertos al azar, es 0 para los ítems 1, 2 y 5 y 0,10 para el 3 y el 4.

En suma, la CCI, piedra angular de la TRI, establece una relación funcional entre los valores de la variable medida θ y la probabilidad de acertar el ítem. El tipo de función matemática adoptado para la CCI, el número de parámetros considerados y otros criterios que se elijan darán lugar a distintos modelos de TRI.

2.2. Unidimensionalidad e independencia local

Como se acaba de señalar en el apartado anterior, la CCI establece una relación funcional entre la probabilidad de acertar un ítem y los valores de θ. Por tanto, si el modelo es correcto, la probabilidad de acertar un ítem *únicamente* dependerá de un factor, a saber, de θ. En otras palabras, la TRI asume implícitamente en su formulación que los ítems destinados a medir la variable θ constituyen una sola dimensión, son unidimensionales. Aunque también se han desarrollado modelos multidimensionales, todo lo dicho hasta ahora se refiere a modelos unidimensionales. En el caso de modelos multidimensionales (MIRT), la CCI recibe la denominación más general de «función característica del ítem», pues ya no es una curva en el plano, sino una función, la que sea, en un espacio multidimensional. Algunas propuestas de modelos multidimensionales pueden verse en Bock y Aitkin (1981), Samejima (1974), Thissen y Steinberg (1984), Whitely (1980), Ackerman (2005), Reckase (2009), o en español el trabajo de Maydeu (1996). En el apartado 4.1 se comentan las estrategias a seguir para comprobar que los datos son unidimensionales y, por tanto, se pueden aplicar los modelos de TRI que asumen este supuesto.

Si se cumple la unidimensionalidad, de ello se deriva que existe *independencia local* entre los ítems; esto es, que para una persona con un determinado valor en la variable unidimensional su respuesta a un ítem no viene influida por sus respuestas en los otros. Nótese que si ello ocurriera, se caería en una contradicción, pues significaría que la variable unidimensional no daría cuenta de toda la varianza de los ítems, sino que parte de esta dependería de otros ítems. La independencia local puede expresarse diciendo que la probabilidad de que una persona

acierte n ítems es igual al producto de las probabilidades de acertar cada uno de ellos. Por ejemplo, si un test consta de tres ítems y la probabilidad de que cierta persona acierte el primero es $P(A_1) = 0{,}40$, de que acierte el segundo $P(A_2) = 0{,}50$ y el tercero $P(A_3) = 0{,}80$, lo que establece el principio de independencia local es que la probabilidad de que esta persona acierte los tres ítems viene dada por:

$$P(A_1, A_2, A_3) = (0{,}40)(0{,}50)(0{,}80) = 0{,}16$$

La probabilidad de que acierte los dos primeros y falle el tercero vendría dada por $p(A_1, A_2, F_3) = = (0{,}40)(0{,}50)(1 - 0{,}80) = 0{,}04$, etc.

Nótese que, por ejemplo, ítems encadenados cuya respuesta a uno depende de que se conozca la del anterior carecerían de independencia local; lo mismo ocurriría si el orden de presentación de los ítems afectase a las respuestas, etc., existiendo distintas situaciones atentatorias contra el principio de independencia local. Análogamente, puede hablarse de independencia local de las personas en el sentido de que el rendimiento de una persona es independiente del rendimiento de las otras.

Matemáticamente, lo dicho anteriormente puede expresarse como sigue:

$$P(U_1, U_2, \ldots, U_n|\theta) = P(U_1|\theta)P(U_2|\theta)\cdots P(U_n|\theta) = \prod_{i=1}^{n} P(U_i|\theta) \qquad [7.3]$$

dónde U_i es la respuesta de una persona al ítem i.

La fórmula indica que, para un valor dado de θ, la probabilidad de un determinado patrón de respuesta a los ítems es igual al producto de las probabilidades para cada uno de los ítems. Esta conceptualización de independencia local es a veces mal interpretada, cuando se considera que la existencia de correlación entre los ítems de un test significa que no hay independencia local. En realidad son cosas distintas: las correlaciones entre las respuestas de las personas a los ítems se deben a su nivel de aptitud, a θ, pero si el valor de θ se parcializa, se mantiene constante, entonces no tiene por qué haber correlación, y este es el caso de la independencia local; nótese que se refiere a un determinado valor de θ. Si hay unidimensionalidad, se obtiene independencia local; en ese sentido son conceptos equivalentes, pero puede existir independencia local con datos no unidimensionales, como es el caso de los modelos multidimensionales de TRI.

EJERCICIOS

1. En la columna (X) de la tabla adjunta aparecen las puntuaciones obtenidas por una muestra de 100 personas en un test de 10 ítems. En la segunda columna (% aciertos) se refleja el porcentaje de personas que, habiendo obtenido la puntuación que figura en la primera columna, han superado el ítem 6 del test.

X	% aciertos
1	5
2	10
3	20
4	30
5	40
6	50
7	65
8	80
9	90
10	95

1. Trace la gráfica de la regresión ítem-test para los datos de la tabla.
2. Según los datos del gráfico anterior, ¿existe alguna conexión entre la puntuación obtenida en el test y la probabilidad de superar el ítem 6? Descríbala.
3. Dibuje un gráfico, correspondiente a un test e ítem hipotéticos, en el que no haya ninguna relación entre las puntuaciones que las personas obtienen en el test y su probabilidad de superar el ítem.
4. ¿Cuál sería la discriminación de un ítem como el del apartado anterior?
5. Señale la diferencia fundamental entre la regresión ítem-test y la curva característica del ítem.

2. En la tabla adjunta aparecen las puntuaciones de una muestra de personas (X) y sus puntua-

ciones en uno de los ítems del test (1 significa acierto y 0 error).

X	% aciertos
4	0
6	1
7	0
8	1
10	1

1. Calcular el índice de discriminación clásico (r_b).
2. Estimar el valor aproximado que tomaría el parámetro a de este ítem.
3. Estimar el valor del parámetro b.

3. En la tabla aparecen las respuestas de 20 personas, todas ellas con el mismo nivel en la variable medida θ, en dos ítems de un test (el 1 significa acierto y el 0 error).

	Personas																			
Ítem A	1	0	1	0	1	0	1	0	1	1	1	0	0	0	1	1	1	0	0	0
Ítem B	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	1	0	0	0

1. Al nivel de confianza del 95%, ¿puede afirmarse que existe independencia local para estos ítems?
2. Genere un patrón de respuestas para las 20 personas en ambos ítems, de tal guisa que se viole totalmente el supuesto de independencia local entre ambos.

4. Se aplicaron dos escalas de actitudes hacia la religión a una muestra de 500 universitarios. La varianza explicada por cada uno de los cinco factores obtenidos tras realizar un análisis factorial de cada una de las escalas fue la siguiente:

	Escala 1	Escala 2
1	12,0	12,0
2	0,9	9,0
3	0,8	0,8
4	0,2	0,8
5	0,1	0,4
Total	**14,0**	**23,0**

1. ¿Qué porcentaje de la varianza total explica el primer factor en cada una de las escalas?
2. Si se desea utilizar alguno de los modelos de TRI con estas escalas, ¿cuál de las dos cumple mejor el supuesto de unidimensionalidad? Razone la respuesta.

SOLUCIONES

1.2. Al aumentar la puntuación en el test, aumenta la probabilidad de superar el ítem.
3. Línea recta paralela al eje de abscisas (puntuaciones en el test).
4. Nula.
5. La escala del test es distinta de la escala θ.

2.1. 0,77.
2. 1,21.
3. –0,32.

3.1. ($\chi^2 = 0 < 3{,}84$): no significativa, se acepta la independencia local.
2. Muchos patrones posibles; los más extremos serían aquellos cuyas tablas de contingencia tuviesen: diez 1-1 y diez 0-0, o diez 1-0 y diez 0-1.

4.1. Escala 1: 86%; escala 2: 52%.
2. Escala 1.

3. MODELOS

Como se ha visto, el distintivo central de la TRI lo constituye la CCI. Ahora bien, según se adopte para la CCI una función matemática u otra y según se tengan en cuenta uno, dos o tres de los parámetros de los ítems descritos, se generarán diferentes modelos. Se han utilizado con mucha frecuencia dos tipos de funciones matemáticas para la CCI, la función logística y la curva normal acumulada, lo que daría lugar a seis modelos según se contemplen uno, dos o tres parámetros para cada una de estas dos funciones. Dada la mayor tratabilidad matemática de la función logística, en la actualidad los tres modelos por antonomasia de la TRI son el logístico de un parámetro, el logístico de dos parámetros y el logístico de tres parámetros. En los tres casos se asume (aquí) que la respuesta a los ítems es dicotómica, es decir, o se acierta o se falla el ítem, independientemente del número de alternativas que tenga, o que sea de carácter abierto en el que las personas deban generar su propia respuesta, en cuyo caso esta solo se considerará correcta o incorrecta, sin grados intermedios. No obstante, existen en la literatura otros tipos de modelos para respuestas multicategoriales (Nering y Ostini, 2010; Revuelta, Abad y Ponsoda, 2006). Para una clasificación de los modelos, véanse Thissen y Steinberg (1986), Goldstein y Wood (1989) o Van der Linden y Hambleton (1997). Se entiende por ítems multicategoriales aquellos en los que las respuestas no se consideran aciertos o fallos, sino que cada una de las alternativas recibe cierta ponderación según su pertinencia. Ítems continuos serían aquellos en los que la persona evaluada estima cierto nivel en una escala. No es infrecuente forzar ambos tipos de ítems dentro de un formato dicotómico; mediante el establecimiento de algún punto de corte, a veces lo que se pierde en información se gana en operatividad.

El número de modelos pensables, barajando diversos criterios clasificatorios, puede acercarse poco menos que a infinito y corresponde a la evidencia empírica y a la enjundia teórica llevar a cabo la selección. En adelante se expondrán con cierto detalle los modelos básicos, como son los logísticos de uno, dos y tres parámetros.

3.1. Modelo logístico de un parámetro (modelo de Rasch)

El modelo logístico de un parámetro fue formulado originalmente por Rasch (1960), recibiendo posteriormente notable atención, especialmente en la Universidad de Chicago, por Wright et al. (Wright, 1977a y b; Wright y Stone, 1979). Es, sin duda, el modelo más popular de TRI, debido en gran parte a la sencillez emanada de su lógica: la respuesta a un ítem solo depende de la competencia de la persona y de la dificultad del ítem, es decir, de θ y de b, lo cual constituye la esencia de la medición.

Según dicho modelo, la CCI viene dada por la función logística, y el único parámetro de los ítems a tener en cuenta es b, el índice de dificultad.

La función logística es una curva cuya fórmula general viene dada por:

$$y = \frac{e^x}{1 + e^x} \qquad [7.4]$$

Se puede representar de forma gráfica (figura 7.6) dando valores a X.

X	Y
$-\infty$	0,000
–3	0,047
–2	0,119
–1	0,269
0	0,500
1	0,731
2	0,881
3	0,953
∞	1,000

La función logística ha sido tradicionalmente utilizada en biología y medicina para problemas como el crecimiento de las plantas o la propagación de enfermedades. Tiene la ventaja sobre la curva normal de que es más fácil para operar con ella matemáticamente, es «matemáticamente más tratable». Mediante el uso de una constante adicional ($D = 1{,}7$), en la función logística sus valores se aproximan notablemente a los de la curva normal acu-

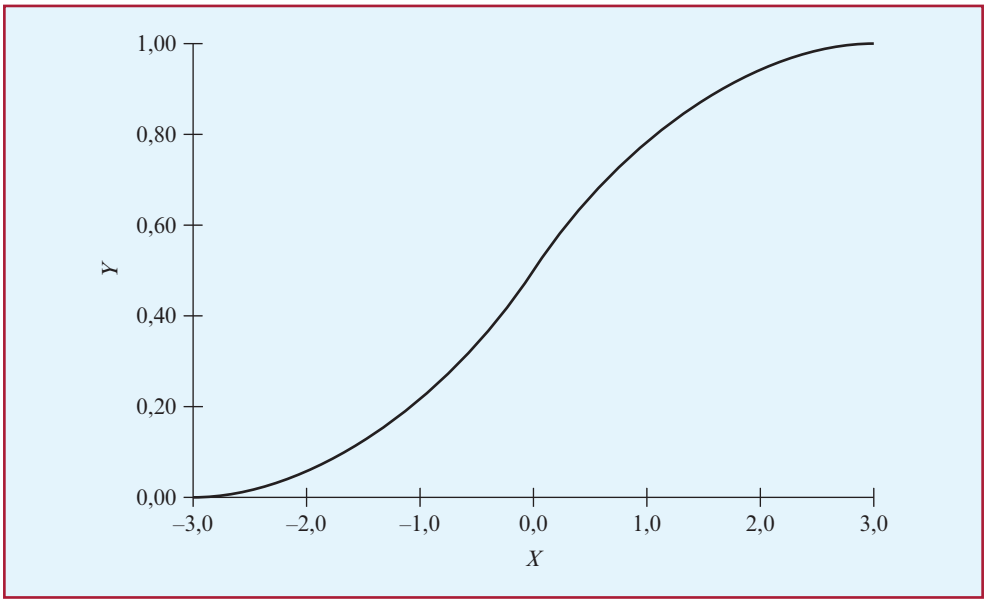


Figura 7.6.—Función logística.

mulada, por lo que es frecuente encontrarla expresada como:

$$y = \frac{e^{DX}}{1 + e^{DX}} \qquad [7.5]$$

Adaptada a la terminología de la TRI para el caso concreto de un parámetro, en el *modelo de Rasch* la CCI adquiere la siguiente expresión:

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \qquad [7.6]$$

donde:

- $P_i(\theta)$: Probabilidad de acertar el ítem i a determinado nivel θ.
- θ: Valores de la variable medida.
- b_i: Índice de dificultad del ítem i.
- e: Base de los logaritmos neperianos (2,72).
- D: Constante (cuando $D = 1{,}7$, los valores de la función logística apenas difieren de los de la curva normal acumulada).

El modelo es claro: conocido el índice de dificultad de un ítem, b, y la competencia de las personas, θ, predice la probabilidad, $P(\theta)$, de que acierten el ítem. (En adelante, para este y otros modelos, cuando el contexto sea inequívoco, se prescindirá del subíndice i, por sencillez.)

Ejemplo

¿Cuál es la probabilidad de que las personas con $\theta = 2$ acierten un ítem cuyo índice de dificultad es $b = 1{,}5$? (Se asume $D = 1$).

$$P(\theta) = \frac{2{,}72^{1(2-1{,}5)}}{1 + 2{,}72^{1(2-1{,}5)}} = 0{,}62$$

Nótese que si $\theta = b$, entonces:

$$P(\theta) = \frac{e^0}{1 + e^0} = \frac{1}{1+1} = \frac{1}{2} = 0{,}5$$

Por tanto, en este modelo puede definirse el índice de dificultad b como aquel valor θ para el cual $P(\theta) = 0{,}5$ (figura 7.7).

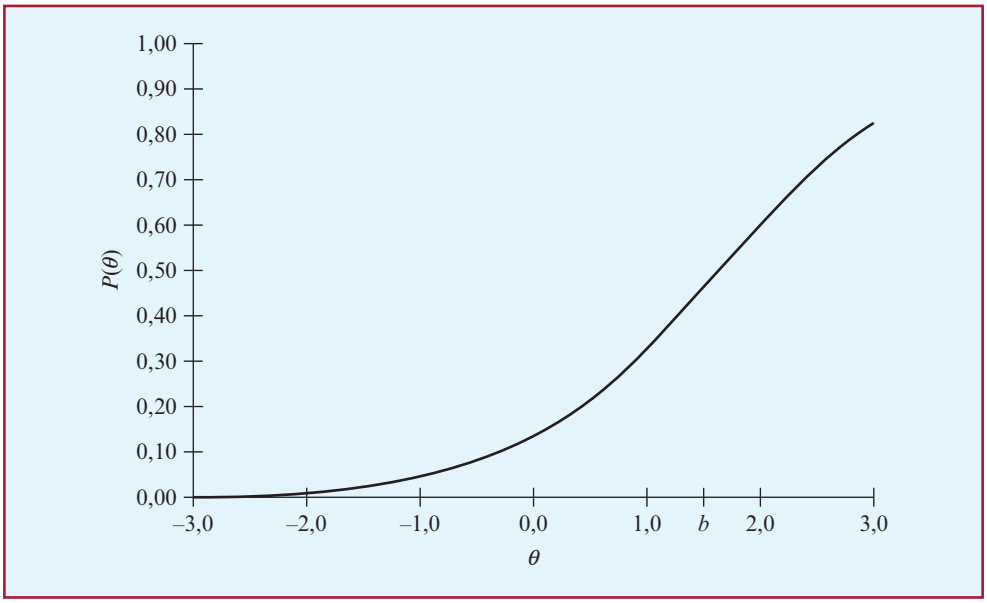


Figura 7.7.—Curva característica de un ítem: $b = 1{,}5$; $c = 0$.

En otras palabras, el índice de dificultad b es el valor de θ en el punto de inflexión de la curva. La ventaja de que b esté en la misma escala que θ será notoria.

Tal vez el lector comience a impacientarse con tanta suposición acerca del valor de θ y de b, cuando en realidad el problema sería cómo hallar su valor. No se olvide que el único dato accesible son las respuestas de las personas a los ítems. A modo de placebo tranquilizador, sepa el lector que, a partir de las respuestas de las personas a los ítems, hay métodos estadísticos razonables implementados en programas informáticos que permiten estimar el valor de b para cada ítem y el valor de θ para cada persona, así como comprobar si el modelo se ajusta a los datos.

Para evitar confusiones, adviértase que la fórmula dada para el modelo de Rasch puede expresarse de formas ligeramente distintas si se hacen algunas operaciones. Así, no es infrecuente encontrarlo formulado del siguiente modo:

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}} \qquad [7.7]$$

que es, por supuesto, equivalente a la fórmula anterior y que viene de dividir su numerador y denominador por $e^{D(\theta - b_i)}$. A su vez, esta última se puede expresar así:

$$P_i(\theta) = [1 + e^{-D(\theta - b_i)}]^{-1} \qquad [7.8]$$

En otras ocasiones, en vez de θ se emplea *exp*, refiriéndose a exponencial. Todas las expresiones son equivalentes, pero aquí se usará la primera citada.

3.2. Modelo logístico de dos parámetros

El modelo logístico de dos parámetros fue originalmente desarrollado por Birnbaum (1957, 1958a, 1958b, 1968). Asume que la CCI viene dada por la función logística y contempla dos parámetros de los ítems, el índice de dificultad b y el índice de discriminación a. Su fórmula viene dada por

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad [7.9]$$

donde, como en el modelo logístico de un parámetro:

$P_i(\theta)$: Probabilidad de acertar el ítem i para un valor θ.

θ: Valores de la variable medida.
b_i: Índice de dificultad del ítem *i*.
a_i: Índice de discriminación del ítem *i*.
e: Base de los logaritmos neperianos (2,72).
D: Constante. Cuando toma el valor 1,7, la función logística se aproxima a la normal acumulada.

EJEMPLO

El índice de discriminación de un ítem es 2, y su índice de dificultad, 1,5. ¿Qué probabilidad tienen de acertar ese ítem las personas cuyo nivel de competencia en la variable medida sea 2,5?

Datos: $a = 2$; $b = 1{,}5$; $\theta = 2{,}5$; $D = 1{,}7$;

$$P(\theta) = \frac{2{,}72^{(1{,}7)(2)(2{,}5-1{,}5)}}{1 + 2{,}72^{(1{,}7)(2)(2{,}5-1{,}5)}} = 0{,}967$$

La probabilidad de superar el ítem es muy elevada (0,967), como era de esperar, pues a medida que θ es mayor que b, para un determinado valor de a, $P(\theta)$ aumenta según el modelo logístico, lo cual es razonable, pues a mayor competencia de las personas, mayor probabilidad de superar un ítem dado. Véanse en la figura 7.8 dos ítems con distinta dificultad y distinto índice de discriminación.

El ítem 2 es más difícil que el 1 ($b_2 > b_1$) y su índice de discriminación es mayor ($a_2 > a_1$). No obstante, a pesar del menor índice de discriminación del ítem 1, nótese que para valores de θ en torno a b_1 la capacidad de discriminación del ítem 1 supera la del 2. Por tanto, la elección de un ítem u otro basándose en su capacidad discriminativa habrá de hacerse en función de la zona de θ que se desea discriminar.

3.3. Modelo logístico de tres parámetros

Con sus orígenes en los trabajos de Birnbaum (1957, 1958a, 1958b, 1968), el modelo logístico de tres parámetros es, junto con el de Rasch, uno de los que más atención ha recibido en la literatura psicométrica. El modelo asume que la CCI viene dada por la función logística y añade a los dos parámetros a y b ya citados un tercero, c, relativo a la probabilidad de acertar el ítem al azar cuando no se conoce la respuesta. Más exactamente c_i es el valor de $P_i(\theta)$ cuando $\theta = -\infty$.

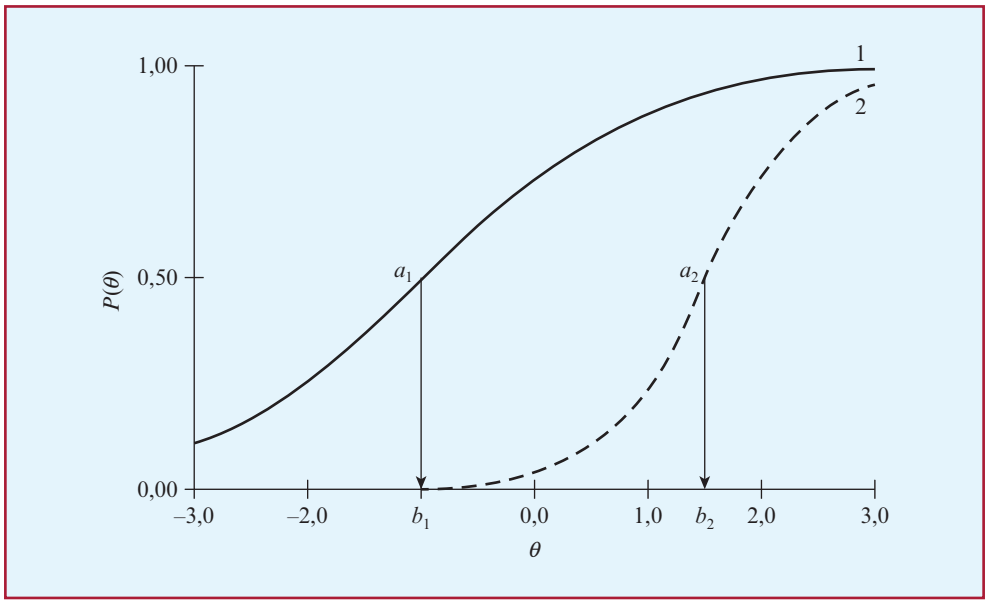


Figura 7.8.—Curvas características de dos ítems con índices de dificultad y discriminación diferentes.

El modelo puede expresarse así:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad [7.10]$$

donde $P_i(\theta)$, e, D, a_i, θ y b_i tienen la misma significación que la ya citada para el caso de dos parámetros, y c_i es el valor de $P_i(\theta)$ cuando $\theta = -\infty$.

Ejemplo

La probabilidad de acertar cierto ítem al azar es 0,25, su índice de dificultad es 0,5 y su índice de discriminación es 1,25. ¿Cuál es la probabilidad de acertar ese ítem para personas con $\theta = 1$?

$$P(\theta) = 0{,}25 + (1 - 0{,}25)\frac{2{,}72^{(1{,}7)(1{,}25)(1-0{,}5)}}{1 + 2{,}72^{(1{,}7)(1{,}25)(1-0{,}5)}} =$$

$$= 0{,}805$$

Este valor de $P(\theta) = 0{,}805$ sería menor si en las mismas condiciones el valor de c fuese cero, en cuyo caso $P(\theta) = 0{,}74$. Naturalmente, con la probabilidad de acertar el ítem al azar de 0,25 la probabilidad de respuestas correctas por parte de las personas aumenta.

Nótese que en los modelos de uno y dos parámetros, cuando el valor de $\theta = b$, $P(\theta) = 0{,}50$, mientras que aquí cuando $\theta = b$, $P(\theta) = (1 + c)/2$. Para los datos del ejemplo, cuando $\theta = b = 0{,}5$, $P(\theta) = (1 + 0{,}25)/2 = = 0{,}625$ (figura 7.9). Ciertamente, si $c = 0$, $P(\theta) = = 1/2 = 0{,}5$; con $c = 0$, el modelo de tres parámetros se convierte en un modelo de dos parámetros.

El modelo logístico de tres parámetros es el más general: si se hace $c = 0$, se obtiene el de dos parámetros, y si además a se asume constante para todos los ítems, se obtiene el de un parámetro.

Algunos autores (Barton y Lord, 1981) han propuesto, incluso, un modelo logístico de cuatro parámetros para tratar de mitigar el problema real de que a veces por determinadas circunstancias, como el descuido o el uso de información que el constructor del ítem no tuvo en cuenta, las personas de alta competencia fallan ítems impropiamente (Hambleton y Swaminathan, 1985). Hasta la fecha se han dedicado pocas investigaciones a este tipo de modelos, y no parece que aporte ventajas significativas respecto al de tres parámetros, máxime cuando los problemas que trata de solucionar más bien hay que evitar que se produzcan. El modelo viene dado por:

$$P_i(\theta) = c_i + (Y_i - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad [7.11]$$

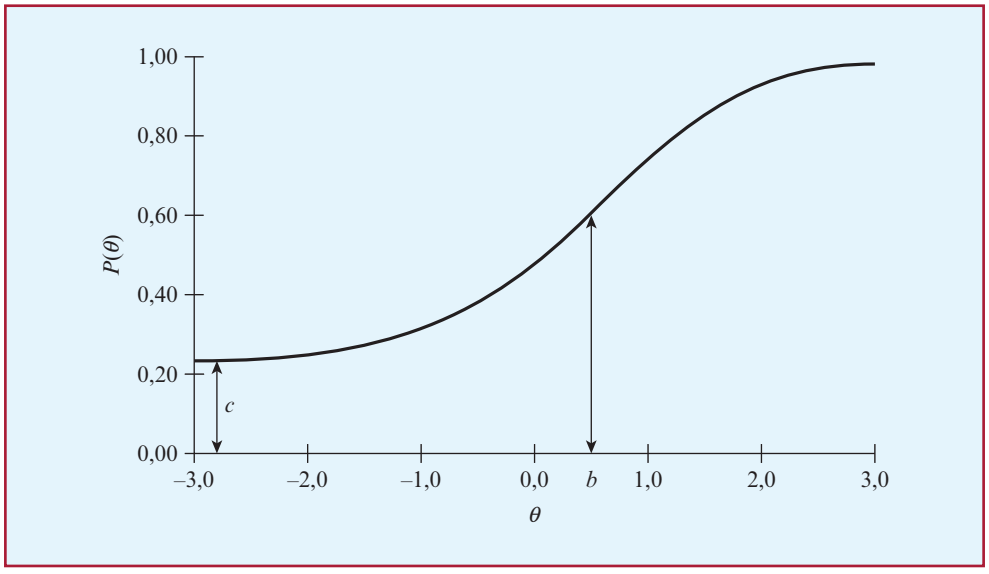


Figura 7.9.—Curva característica de un ítem con $b = 0{,}5$ y $c = 0{,}25$.

Fórmulas de los modelos logísticos

Un parámetro: $$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$

Dos parámetros: $$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Tres parámetros: $$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Cuatro parámetros: $$P_i(\theta) = c_i + (Y_i - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

donde Y_i toma valores ligeramente inferiores a 1 y el resto de los componentes son los ya descritos para los otros modelos.

3.4. Modelos de ojiva normal

Los modelos de ojiva normal asumen que la CCI viene dada por la función de la curva normal acumulada. Preceden en su desarrollo a los modelos logísticos (Lawley, 1943; Tucker, 1946; Lord, 1952), pero, como ya se ha señalado, la mayor tratabilidad matemática de la función logística ha determinado su predominio. Como señala Lord (1980), no hay razones sustantivas sólidas para elegir a priori un tipo de modelo u otro (logístico/normal), y a nivel práctico los resultados son muy similares, por lo que la mayor manejabilidad matemática decide a inclinarse por los logísticos. Otra razón apuntada por Lord para esta preferencia es que, aunque teóricamente los personas con una elevada competencia no deberían fallar ítems fáciles, por razones varias, de hecho lo hacen, y como la función logística se aproxima asintóticamente más lentamente que la normal, este tipo de anomalías tendrán menos incidencia sobre el modelo logístico que sobre el normal.

Aquí se expondrán las fórmulas de los modelos de ojiva normal con carácter ilustrativo y teórico, pero por las razones expuestas su utilización actual es escasa. Nótese que hasta su pionero y exponente máximo (Lord, 1952) se «convirtió» a la fe logística (Lord, 1968, 1980), parece que bajo las influencias de Birnbaum (Wright y Stone, 1979). En el cuadro adjunto aparecen formulados los modelos de ojiva normal de uno, dos, tres y cuatro parámetros, donde θ, a, b, c y e tienen la misma significación que la ya vista en los modelos logísticos.

Modelos de ojiva normal

Un parámetro: $$P_i(\theta) = \int_{-\infty}^{\theta - b_i} (1/\sqrt{2\pi})e^{(-Z^2/2)}\,dz$$

Dos parámetros: $$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} (1/\sqrt{2\pi})e^{(-Z^2/2)}\,dz$$

Tres parámetros: $$P_i(\theta) = c_i + (1 - c_i)\int_{-\infty}^{a_i(\theta - b_i)} (1/\sqrt{2\pi})e^{(-Z^2/2)}\,dz$$

Cuatro parámetros: $$P_i(\theta) = c_i + (Y_i - c_i)\int_{-\infty}^{a_i(\theta - b_i)} (1/\sqrt{2\pi})e^{(-Z^2/2)}\,dz$$

3.5. Orígenes y desarrollo de la TRI

A continuación se ofrece una breve panorámica histórica del nacimiento y evolución de los modelos de TRI; para una revisión más amplia puede consultarse el trabajo de Muñiz y Hambleton (1992), en el cual se basan las líneas que siguen, o el más reciente de Faulkner y Wells (2016). Los trabajos pioneros, que ahora retrospectivamente pueden verse como el germen de lo que posteriormente ha dado en llamarse TRI, se deben una vez más al genio de Thurstone (Thurstone, 1925, 1928a, 1928b; Thurstone y Ackerson, 1929). En especial, el trabajo de 1925 podría considerarse un claro antecedente de las curvas características de los ítems, cuando Thurstone presenta una serie de curvas conectando la edad de las personas con la proporción de aciertos de cada ítem, tomados del test de inteligencia de Binet. Tucker (1987), en su revisión de los métodos clásicos de análisis de ítems, señala también este trabajo como uno de los pioneros, y se atribuye de paso haber acuñado por primera vez hacia 1945 el término «curva característica del ítem», acuñación que reconoce Lord (1952).

Al lado de estos primeros atisbos, hay que citar a Binet y Simon (1905a, 1905b, 1908), cuyos gráficos de la evolución de los niños según la edad pueden considerarse una primera aproximación a curvas características rudimentarias. Asimismo, el trabajo de Richardson (1936) es seguramente el primer intento de ajustar la ojiva normal a las respuestas a los ítems. Sus consejos acerca de la necesidad de controlar la dificultad de los ítems, en función de los objetivos perseguidos por el test, representan una formulación verbal anticipada de lo que luego habría de permitir realizar la función de información en el marco de la TRI. Ferguson (1942) también se acerca, vía los métodos psicofísicos, al planteamiento de las curvas características de los ítems. El paralelismo de tratar las proporciones de aciertos en los ítems frente a los valores globales en el test, en los mismos términos en que lo venían haciendo los métodos psicofísicos para la determinación de los umbrales, será una característica común en estos comienzos. El propio Ferguson (1942) señala explícitamente que en los últimos años se da una tendencia creciente entre los psicómetras a acercar sus métodos a los de la psicofísica. Nada más natural que, a la hora de determinar los parámetros de los ítems, los teóricos de los test acudiesen a los métodos psicofísicos clásicos, en concreto al de los estímulos constantes, pues tenían el mismo problema que estos para determinar el umbral absoluto, aquel valor en el eje de abscisas detectado el 50% de las veces, para lo cual se estaba utilizando la función psicométrica bajo la hipótesis phi-gamma (Blanco, 1996; Muñiz, 1991), conceptualmente equivalente al parámetro b (dificultad), o valor de θ cuando $P(\theta) = 0{,}50$, supuesto que no hay aciertos al azar.

Lawley (1943, 1944) lleva a cabo una aproximación más sistemática para modelos muy restrictivos, y Tucker (1946) también utiliza la curva normal como rudimento de curva característica. Suele atribuirse a Lazarsfeld (1950) la paternidad del término «rasgo latente», que será el nombre que tomarán en principio los modelos, aunque posteriormente se haya generalizado el de TRI, pues refleja mejor su funcionamiento real, basado en los ítems, permitiendo además distinguirlos de otras modelizaciones que también utilizan el término «latente», tales como el análisis factorial, ecuaciones estructurales o análisis multidimensional (Hambleton y Swaminathan, 1985). En su revisión de la TRI, Goldstein y Wood (1989) proponen que el término «teoría» se cambie por modelos, ya que más que teorías psicológicas explicativas lo que se hace es modelizar las respuestas a los ítems, pero a estas alturas el término TRI aparece consolidado.

Si bien estos pueden ser considerados los orígenes remotos, el nacimiento formal podría ubicarse en los trabajos de Lord (1952, 1953a, 1953b), que representan la semilla de la que saldrán los frutos de la TRI actual. El trabajo de Lord (1952) es el resultado de su tesis doctoral, dirigida por Gulliksen y asesorado por Tucker, la flor y nata psicométrica de la época. Representa junto con sus propios trabajos de 1953 (Lord, 1953a, 1953b) la formulación más sistemática de los principales conceptos de la TRI, a partir de los cuales surgirán los desarrollos posteriores. Si hubiera que ubicar puntualmente en algún momento los orígenes genuinos de la TRI, lo haríamos en estos trabajos de Lord, especialmente en el de 1952. Este enfoque marcará un rumbo diferente en las investigaciones psicométricas, si bien, como el propio Lord indica, las conclusiones obtenidas bajo la nueva óptica no contradicen en general los grandes logros de la teoría clásica de los test, sino que más bien los complementan.

Nace un nuevo enfoque, pero por entonces aún faltan 30 años para que los modelos de TRI dominen el escenario psicométrico. Birnbaum (1957, 1958a, 1958b) da otro gran empujón al campo, sustituyendo los modelos de ojiva normal de Lord por los logísticos, más tratables matemáticamente, generando los desarrollos matemáticos necesarios para su posible y futuro uso aplicado. En 1960 el danés George Rasch publica un famoso libro en el que expone con detalle el modelo logístico de un parámetro, utilizando material de test de aptitudes. Rasch es consciente de que su trabajo supone un cambio radical en el enfoque psicométrico, y en la introducción expone con claridad cómo su modelo viene a resolver los problemas de invarianza previamente mencionados.

Nótese que para estas fechas nada de lo dicho y hecho se traduce en una aplicabilidad directa y generalizada de los modelos por parte de los posibles usuarios, nos movemos a nivel teórico matemático. El impulso más potente llegará con la publicación en 1968 del libro de Lord y Novick en el que se dedican cinco capítulos al tema, cuatro de ellos escritos por Birnbaum. Llegados ahí, 1968, puede decirse que el grueso del corpus general está escrito, y los primeros modelos formulados, pero su implantación y progreso serán lentos y laboriosos debido a la complejidad matemática de los modelos, a la ausencia de programas informáticos disponibles para analizar los datos según los nuevos modelos y al escepticismo general acerca de las ventajas de esta nueva línea de investigación (Hambleton y Swaminathan, 1985). En su trabajo de 1969 Wright y Panchapakesan desarrollan la estimación de los parámetros para el modelo de Rasch, introduciendo el programa BICAL. Bock y Wood (1971) incluyen por primera vez en las revisiones para el Annual Review un apartado dedicado a la entonces denominada «teoría del rasgo latente», donde exponen con claridad las ventajas de los nuevos modelos y la literatura sobre el tema por entonces, haciendo especial hincapié en el libro de Lord y Novick (1968) y en el de Rasch (1960), como no podía ser de otro modo.

Entre la publicación del libro de Lord y Novick (1968) y la disponibilidad de los principales programas informáticos, BICAL (Wright y Mead, 1976; Wright, Mead y Bell, 1979), LOGIST (Wood, Wingersky y Lord, 1976; Wingersky, Barton y Lord, 1982), BILOG (Mislevy y Bock, 1984), MULTILOG (Thissen, 1986), MICROCAT (Assessment Systems Corporation, 1988), NOHARM (Fraser y McDonald, 1988), ANCILLES y OGIVA (Urry, 1977), fundamentales para poder aplicar los modelos, transcurre una década de rápido crecimiento de la literatura y los avances en TRI, empezando a vislumbrarse con claridad las posibilidades reales de la aplicación práctica de los modelos. Especial mención por su militancia en pro de la TRI merecen el profesor Benjamin Wright y su grupo de Chicago; su conferencia invitada de 1967 (Wright, 1968) en un congreso organizado por el Educational Testing Service (ETS) en Nueva York, sobre los problemas de los test, suele tenerse por muy influyente, estimulando a los constructores de test a utilizar la nueva tecnología de la TRI entonces emergente. Samejima (1969) extiende los modelos para ítems de respuesta no dicotómica, Bock (1972) y Lord (1974) proponen nuevos métodos de estimación de los parámetros, para lo cual Lord utiliza el programa LOGIST.

El desarrollo teórico es rápido, pero será la década de los ochenta-noventa la que supondrá la verdadera expansión y afianzamiento de la TRI y su masivo predominio en psicometría. El punto de inflexión puede ubicarse en otro libro, cómo no, de Frederic Lord publicado en 1980 y sintomáticamente titulado *Aplicaciones de la Teoría de Respuesta a los Ítems a problemas prácticos del uso de los test,* pues, efectivamente, las aplicaciones habían llegado. En este excelente libro, hito bibliográfico de la TRI, Lord recoge tanto los desarrollos teóricos como las aplicaciones de los modelos de TRI disponibles hasta entonces. A partir de esas fechas, los trabajos sobre distintos aspectos monográficos de la TRI se multiplican y las revistas del área y los congresos así lo reflejan; finalmente, el enfoque de la TRI es dominante en el ámbito psicométrico, y un nuevo paradigma psicométrico pasa a dominar la escena. En 1982 la revista *Applied Psychological Measurement* dedica un número monográfico a la TRI, y aparece toda una serie de trabajos que cubren los distintos aspectos que se han ido desarrollando durante los años anteriores, entre los que cabe citar los de Hulin, Drasgow y Parsons (1983), Hambleton y Swaminathan (1985), Baker (1985), Andrich (1988), Linn (1989), Hambleton et al. (1991), Van der Linden y Hambleton (1997), Bock (1997), Ayala (2009),

Wells y Faulkner-Bond (2016), entre otros muchos. Para una bibliografía exhaustiva clasificada sobre la TRI, véase Hambleton (1990).

A continuación se presenta una cronología con algunos de los hitos más destacados en el desarrollo de la TRI.

Cronología de los modelos de TRI

1905	Binet y Simon anticipan con sus gráficos el concepto de curva característica.
1925	Thurstone presenta curvas semejantes a las curvas características de los ítems.
1936	Richardson ajusta la ojiva normal a las respuestas a los ítems.
1942	A través de los métodos psicofísicos, Ferguson se aproxima al planteamiento de las curvas características.
1944	Guttman presenta su modelo de escalamiento para datos cualitativos.
1945	Tucker acuña el término «curva característica del ítem».
1946	Lazarsfeld acuña el término de «rasgo latente», denominación inicial de los modelos de TRI.
1952	Lord publica *Una teoría sobre las puntuaciones de los test,* que puede ser considerado el inicio de la TRI.
1957	Bimbaum utiliza la función logística en vez de la ojiva normal.
1960	Publicación del libro de Rasch sobre el modelo logístico de un parámetro.
1967	Influyente conferencia invitada de Benjamin Wright en un congreso en Nueva York organizado por el Educational Testing Service.
1968	Se publica el libro de Lord y Novick, con contribuciones de Birnbaum, que supone un fuerte impulso para la TRI.
1969	Wright y Panchapakesan publican su trabajo sobre la estimación de los parámetros del modelo de Rasch e introducen el programa BICAL. Samejima propone modelos para ítems politómicos y de respuesta continua.
1972	Nuevos desarrollos en la estimación de los parámetros.
1974	Lord presenta sus nuevos métodos de estimación, implementados en el programa de ordenador LOGIST.
1976	Se hace público el programa LOGIST para estimar los parámetros de los modelos logísticos.
1979	Libro de Wright y Stone, *Best Test Design.* Programa de ordenador BICAL para el modelo logístico de un parámetro.
1980	Lord publica su libro sobre aplicaciones de la TRI.
1981	Bock y Aitkin presentan el método de estimación de máxima verosimilitud marginal, incorporado en el programa BILOG.
1982	Segunda versión del programa de ordenador LOGIST para estimar los parámetros de los modelos logísticos.
1984	Programa de ordenador BILOG para estimar los parámetros de los modelos logísticos.
1985	Libro de Hambleton y Swaminathan.
1990	Libro de Wainer sobre los test adaptativos computerizados.
1991	Libro de Hambleton, Swaminathan y Rogers, seguramente el texto introductorio sobre TRI más utilizado en la actualidad.
1993	Holland y Wainer coordinan un libro sobre el funcionamiento diferencial de los ítems.
1997	Manual de TRI coordinado por Van der Linden y Hambleton.
2005	Libro en honor de Roc McDonald, editado por McArdle y Maydeu.
2009	Libro de Reckase sobre modelos multidimensionales (MIRT).
2010	Manual editado por Nering y Ostini sobre modelos politómicos.
2013	Software flexMIRT para modelos multidimensionales (Cai, 2013).
2016	Libro en honor a Ronald K. Hambleton, uno de los pioneros de la TRI, editado por Wells y Faulkner-Bond.
2016	Manual editado por Van der Linden sobre modelos, métodos y aplicaciones de la TRI.

Algunas diferencias entre la teoría clásica y la TRI

Se presenta a continuación un cuadro con algunos de los aspectos fundamentales en los que se diferencian la teoría clásica y la TRI.

Aspectos	Teoría clásica	Teoría de respuesta a los ítems
Modelo.	Lineal	No lineal.
Asunciones.	Débiles (fáciles de cumplir por los datos).	Fuertes (difíciles de cumplir por los datos).
Invarianza de las mediciones.	No.	Sí.
Invarianza de las propiedades del test.	No.	Sí.
Escala de las puntuaciones.	Entre cero y la puntuación máxima en el test (o alguna transformación de estas).	Entre $-\infty$ y $+\infty$ (o alguna transformación de estas).
Énfasis.	Test.	Ítems.
Relación ítem-test.	Sin especificar.	Curva característica del ítem.
Descripción de los ítems.	Índice de dificultad. Índice de discriminación.	Parámetros *a*, *b*, *c*.
Errores de medida.	Error típico de medida (común para toda la muestra).	Función de información (varía según el nivel en la variable medida).
Tamaño muestral.	Puede funcionar bien con muestras entre 200 y 500 personas, aproximadamente.	Se recomiendan más de 500 personas, aunque depende del modelo.

A continuación se comenta brevemente cada uno de los aspectos diferenciales mencionados en el cuadro.

Modelo

En la teoría clásica el modelo utilizado es lineal, la puntuación empírica es igual a la verdadera más el error ($X = V + e$), mientras que en la TRI la función que relaciona las puntuaciones empíricas con las verdaderas es curvilínea, viene dada por el tipo de curva adoptada por el modelo, habitualmente logística, aunque otras muchas son posibles.

Asunciones

Las asunciones del modelo clásico son débiles en el sentido de que son generales y es fácil que la mayoría de los datos empíricos las cumplan; su fuerza está en su generalidad, pues son aplicables a situaciones muy variadas. Por el contrario, las asunciones de la TRI son más fuertes, más restrictivas, se sacrifica la generalidad para ganar precisión predictiva. El precio a pagar es la exigencia de que los datos cumplan supuestos muy específicos. Nótese que el modelo lineal clásico original, denominado «modelo clásico débil», no hace ninguna asunción sobre las distribuciones de las puntuaciones ni de los errores; cuando se exige que los errores se distribuyan según la curva normal, suele hablarse de modelo clásico fuerte, pero, así y todo, las asunciones son suaves comparadas con las de la TRI. Estas exigencias mínimas del modelo clásico constituyen a la vez su fuerza y su debilidad; por un lado, permiten su uso en un abanico muy amplio de situaciones empíricas, lo cual está muy bien, pero, por contra, las predicciones resultan más genéricas. Ante la eterna disyuntiva entre generalidad y precisión, a la que toda metodología científica se enfrenta, la teoría clásica da más peso a la generalidad y la TRI a la precisión. Ambos enfoques están condenados a entenderse en provecho de los usuarios.

Invarianza de las mediciones

El punto fuerte de la TRI frente al modelo clásico está en que permite mediciones invariantes respecto del instrumento utilizado, propiedad clave en

toda medición. La teoría clásica sobrellevó este déficit en la práctica de forma digna, elaborando toda una tecnología para equiparar las puntuaciones obtenidas con distintos instrumentos. Una buena exposición de la tecnología utilizada puede consultarse en Navas (1996). La invarianza de las mediciones en la TRI se deriva de los modelos utilizados, pero ello no exime de su comprobación empírica, para lo cual se pueden utilizar distintas estrategias, como se verá más adelante.

Invarianza de las propiedades del test

Propiedades tan importantes de un test como el índice de dificultad de los ítems, o su índice de discriminación, en la teoría clásica dependen de la muestra de personas utilizadas para estimarlas. Por ejemplo, si las personas tienen un nivel bajo en la variable medida, la dificultad de los ítems resultará elevada; por el contrario, si el nivel de la muestra es alto, la dificultad de esos mismos ítems será baja. Es decir, un mismo ítem tendría distinto índice de dificultad en función de la muestra utilizada para calcularlo. Los usuarios de la teoría clásica disponen de una solución práctica a este problema, consistente en estimar las propiedades del test en muestras de personas extraídas de la población con la que se va a usar, y no extender sus propiedades más allá de esa población. En el caso del índice de dificultad, no se hablará del índice en general, sino del índice para determinada población. Este acercamiento es perfectamente legítimo y correcto en la práctica, pero poco satisfactorio desde el punto de vista teórico, pues implica asumir tantos valores para las propiedades del instrumento como posibles poblaciones de personas con las que se utiliza.

Como ocurría con las mediciones, la invarianza de las propiedades de los instrumentos se deriva de los modelos de TRI, pero ha de comprobarse empíricamente. Un error muy común es pensar que estas invarianzas se dan por arte de magia, nada más lejano de la realidad empírica; como ocurre con cualquier otro procedimiento de estimación estadística, cuanto mayores sean la amplitud y variabilidad de las muestras utilizadas, mayores serán la precisión con la que se estiman los parámetros de la TRI y, por ende, su invarianza.

Escala de las puntuaciones

Como es bien sabido, en la teoría clásica la escala empírica de las puntuaciones va desde la puntuación mínima obtenible en el test, habitualmente cero, hasta la máxima puntuación posible. No obstante, para facilitar la interpretabilidad de las puntuaciones, esta escala suele transformarse en otra más comprensible, o más conveniente por las razones que sean, por ejemplo, percentiles, puntuaciones típicas, decatipos, cocientes intelectuales, eneatipos, etc. En la TRI también se hacen estas transformaciones de conveniencia, pero la diferencia clave es que las puntuaciones estimadas a las personas (θ) van de menos infinito a más infinito, y en esa escala aparecen todas las mediciones, se use el test que se use; de ahí la mentada invarianza. Esto es muy contraintuitivo para las personas no familiarizadas con la TRI, pues es difícil de entender que si se aplica un test de, pongamos, 40 ítems, las puntuaciones obtenidas por las personas estén entre $-\infty$ y $+\infty$. En la práctica esto no supone mayor problema, puesto que las puntuaciones a los usuarios y clientes se ofrecen transformadas en escalas fáciles de comprender.

La función que une las puntuaciones θ con las puntuaciones en la escala del test se denomina «curva característica del test».

Énfasis

El propio nombre de teoría de respuesta a los ítems ya alude a que, bajo la óptica de la TRI, la unidad de análisis básica es el ítem y no el test, como ocurría en la teoría clásica. El test pasa a ser un agregado de ítems y sus propiedades dependen de las de estos. Puesto que cualquier agregado de ítems (test) que se elija proporciona una medición en la misma escala común, cuál de los posibles test se utilice de los muchos que se pueden escoger a partir de un banco de ítems deja de ser esencial, pues los resultados son igualmente comparables. Esto no era así en la teoría clásica, en la cual, para poder comparar las mediciones de dos personas, había que aplicarles el mismo test o dos formas paralelas. Por ejemplo, en el caso de los test adaptativos informatizados, a cada persona se le aplica un test distinto, dependiendo de su competencia en la variable medida, lo cual sería inconcebible desde el

punto de vista clásico, en el cual al cambiar el test cambiaría la escala, obligando a un tedioso proceso de equiparación de las puntuaciones obtenidas con distintos test para la misma variable.

Relación ítem-test

En la teoría clásica, aunque sepamos la puntuación de una persona en un test, no por ello conocemos la probabilidad que tiene de acertar determinado ítem del test; el modelo no establece una conexión formal entre la puntuación en el test y la probabilidad de superar los ítems. Por el contrario, en la TRI, una vez definida la curva característica del ítem, si conocemos la puntuación de una persona es inmediato el cálculo de la probabilidad que tiene de superar el ítem, es decir, la CCI conecta las puntuaciones de las personas con las probabilidades de superar el ítem. Sin duda esta propiedad es una clara ventaja de la TRI sobre el enfoque clásico, que tendrá consecuencias muy deseables para la construcción, análisis y uso de los test.

Descripción de los ítems

Los dos índices utilizados más frecuentemente para describir los ítems dentro del marco clásico son el índice de dificultad y el de discriminación. Ambos resultan dependientes de la muestra en la que se calculan, mientras que los parámetros *a*, *b* y *c* utilizados en la TRI son invariantes respecto de la muestra utilizada para estimarlos. Una propiedad del parámetro de dificultad *b*, que tendrá ventajas muy notables, es que viene expresado en la misma métrica que las puntuaciones θ de las personas.

Errores de medida

Una de las grandes ventajas que siempre ha reclamado la TRI frente al modelo lineal clásico es la de ser capaz de ofrecer errores de medida en función del nivel de las personas en la variable medida, valiéndose de la función de información. Ello no deja de ser cierto, pues lo habitual en la teoría clásica es ofrecer un error único y común, el error típico de medida para todas las personas de la muestra, sin tener en cuenta su nivel en la variable medida. Pero también es verdad que ya Thorndike (1951) propuso el cálculo del error típico de medida para distintos niveles de competencia. Esta línea ha seguido progresando, impulsada sobre todo por el grupo de la Universidad de Iowa, y en la actualidad se dispone de todo un conjunto de métodos refinados para estimar el error típico de medida a distintos niveles de la variable medida (Feldt et al., 1985; Feldt y Qualls, 1996; Lord, 1984; Qualls, 1992). Justo es reconocer que el enfoque clásico también sabía cómo tratar con los errores para distintos niveles de la variable medida, pero la elegancia conceptual y formal de la función de información de la TRI supera con creces en este punto al enfoque clásico.

Tamaño muestral

Para estimar con precisión los parámetros de los ítems y la puntuación de las personas, la TRI necesita muestras muy amplias, puede decirse que a partir de 500 personas, aunque es un número orientativo. Esto no supone mayor problema para las grandes agencias y compañías administradoras de test, pero para los profesionales de a pie puede constituir un inconveniente serio. Aquí el modelo clásico recupera terreno otra vez, pues funciona razonablemente bien con muestras bastante menores, en torno a 200. Conviene, por tanto, señalar de nuevo que la combinación de ambos acercamientos, clásico y TRI, en función de las circunstancias, parece lo más recomendable, como así sucede habitualmente en la práctica.

EJERCICIOS

1. Señale cuáles de las siguientes características corresponden a la teoría clásica (TC) y cuáles a la teoría de respuesta a los ítems (TRI).

1. Establece un modelo lineal.
2. Los datos cumplen fácilmente los supuestos del modelo.

3. Existe invarianza de las mediciones respecto del test utilizado.
4. Las propiedades del test predominan sobre las de los ítems.
5. Establece el mismo error típico de medida para toda la muestra.
6. Se especifica la relación entre cada ítem y el test.

2. Se aplicó una prueba de inteligencia de cuatro ítems a una muestra de 400 personas. Mediante el programa BILOG se estimaron los valores de los parámetros *a*, *b* y *c* de cada uno de los ítems, obteniendo los siguientes resultados:

Ítems	*a*	*b*	*c*
1	1,6	1,0	0,00
2	0,5	0,0	0,00
3	1,2	1,5	0,20
4	1,0	–0,5	0,25

1. ¿Qué modelo se utilizó para estimar los parámetros de los ítems? Razone la respuesta.
2. ¿Cuál es el ítem más difícil?
3. ¿Cuál es el ítem más fácil?
4. ¿Cuál es el ítem que es más probable que acierte una persona con inteligencia baja que responde al azar?
5. Represente gráficamente las curvas características de los cuatro ítems.
6. Calcule la probabilidad que tienen las personas con una puntuación $\theta = 2$ de superar cada uno de los ítems.
7. Calcule la probabilidad de que las personas con una puntuación $\theta = 2$ superen todos los ítems del test.
8. ¿Con qué ítems funcionaría mejor el modelo logístico de dos parámetros?
9. ¿Cuál de los ítems resultaría más fácil para una persona cuya puntuación en la variable medida θ fuese –0,5? ¿Qué probabilidad tendría esa persona de fallar dicho ítem?
10. ¿Cuál es la probabilidad de que la persona del apartado anterior ($\theta = -0,5$) supere al menos tres de los ítems?

SOLUCIONES

1.1. TC
2. TC
3. TRI
4. TC
5. TC
6. TRI

2.1. 3-p
2. 3
3. 4
4. 4
6. 0,94, 0,85, 0,79, 0,99
7. 0,62
8. 1, 2
9. 4; 0,37
10. 0,06

4. APLICACIÓN DE LOS MODELOS

Para utilizar en la práctica los modelos unidimensionales descritos hay que empezar por comprobar que los ítems constituyen una sola dimensión, tal como exigen los modelos. Una vez comprobado, hay que elegir el modelo a utilizar, luego estimar el valor de parámetros y finalmente comprobar que el modelo estimado se ajusta a los datos empíricos, pues a la postre serán estos el criterio último que nos haga aceptar o rechazar el modelo. Se expone a continuación la lógica de cada uno de estas fases de la aplicación de los modelos.

4.1. Comprobación de la unidimensionalidad

Como ya se ha señalado anteriormente, para que se puedan aplicar los modelos básicos de la TRI los ítems del test deben ser unidimensionales;

por tanto, antes de utilizar uno de estos modelos hay que comprobar la unidimensionalidad. Para comprobar que un conjunto de ítems constituye una sola dimensión existen de antiguo diversas alternativas, habiéndose propuesto numerosos índices al respecto. Hattie (1985), en un buen análisis y clasificación de ellos, da cuenta de ochenta y siete distintos. Estudios comparativos pueden verse en Hambleton y Rovinelli (1986), Hattie (1984) o Zwick y Velicer (1986), y un buen tratamiento puede consultarse en Wells, Rios y Faulkner-Bond (2016), Swaminathan, Hambleton y Rogers (2007), Tate (2003), Svetina y Levy (2014), y en castellano Cuesta (1996).

Veamos algunas posibilidades de comprobar la unidimensionalidad, empezando por el método más clásico: el análisis factorial exploratorio (AFE). Es uno de los métodos tradicionalmente más utilizado, y si bien hoy existen métodos más eficaces, conviene conocer la lógica que subyace a este método que ha teñido la historia de la psicología. Dado que empíricamente raras veces, si alguna, se encuentra una unidimensionalidad perfecta, esto es, que un solo factor dé cuenta de un 100 por 100 de la varianza de los ítems, la unidimensionalidad se convierte en una cuestión de grado, es decir, siempre habrá más o menos unidimensionalidad, por lo que el problema será dónde establecer el punto de corte para asegurar que un conjunto de datos son unidimensionales. Lumsden (1961), por ejemplo, propone como índice de unidimensionalidad el cociente entre la varianza explicada por el primer factor y la explicada por el segundo, pero como bien señala Lord (1980), se necesitan procedimientos estadísticos más rigurosos. Un criterio práctico para decidir sobre la unidimensionalidad podría ser el sugerido por el propio Lord (1980) de extraer las raíces latentes de la matriz de correlaciones tetracóricas entre los ítems, con las comunalidades en la diagonal, y, si la primera raíz latente es «notablemente» superior a la segunda y esta no difiere «mucho» del resto, los ítems pueden considerarse aproximadamente unidimensionales. Por su parte, la varianza explicada por el primer factor es un indicador clásico y muy intuitivo de la unidimensionalidad, pues muestra en qué grado la reducción de datos obtenida (se pasa de *n* ítems a un factor) aún explica una parte importante de la varianza original explicada por los *n* ítems. La pregunta de fondo es ¿cuánta varianza estamos dispuestos a sacrificar para pasar de *n* ítems a un solo factor o dimensión? Para tomar esa decisión hay que apoyarse en dos pilares: por un lado, en la teoría sustantiva que guía nuestro trabajo, y por otro, en criterios estadísticos rigurosos. Es la combinación prudente de estos dos criterios la que debe guiarnos, ninguno de los dos por sí solo es suficiente. Los trabajos de Ferrando y Anguiano (2010), Izquierdo, Olea y Abad (2014), Lloret-Segura, Ferreres-Traver, Hernández-Baeza y Tomás-Marco (2014) nos dan indicaciones muy pertinentes para determinar el número de factores, y por ende para evaluar la unidimensionalidad de los ítems.

En cuanto al software, el programa FACTOR es altamente recomendable por su flexibilidad, facilidad de uso y acceso libre (Ferrando y Lorenzo-Seva, 2017). Por su parte, Deng y Hambleton (2007) revisan nada menos que veinte programas informáticos para evaluar la dimensionalidad de unos datos, cada uno de ellos con sus pros y sus contras, de modo que hay donde elegir. Un aspecto relativamente tranquilizador a la hora de utilizar modelos de TRI con datos no estrictamente unidimensionales es que los estudios de simulación indican que los modelos son robustos a violaciones moderadas de la unidimensionalidad (Ansley y Forsyth, 1985; Cuesta, 1996; Drasgow y Parsons, 1983; Greaud, 1988; Harrison, 1986; Muñiz y Cuesta, 1993; Reckase, 1979; Yen, 1984). En general, como es lógico, el deterioro del funcionamiento de los modelos se acrecienta a medida que se va deteriorando la unidimensionalidad. Una advertencia final: si se utiliza el AFE lineal con datos categóricos, lo cual es muy habitual en la historia de la psicología y otras ciencias, hay que ser muy prudentes a la hora de interpretar los resultados, debido a los llamados «factores de dificultad» que se generan. En esas circunstancias, mejor usar otras alternativas. Siguiendo las sabias palabras de Box y Draper (1987), bien podría decirse aquí que en esencia todos los modelos son erróneos, pero que algunos resultan útiles.

El análisis factorial confirmatorio (AFC) es una opción más aconsejable que el AFE a la hora de evaluar la unidimensionalidad de unos datos (Brown, 2006; Kline, 2015). Tiene ventajas sobre el AFE, entre otras, que permite establecer especificaciones a priori, sobre la estructura dimensional de los datos, basándose en las teorías sustantivas que se manejen y en los resultados previos. Si se somete

a prueba la unidimensionalidad de unos ítems mediante AFC, existen distintos indicadores para comprobar el ajuste obtenido (Ferrando y Anguiano, 2010; Hu y Bentler, 1999) y tomar la decisión al respecto. Nótese que como bien señalan Ferrando y Anguiano (2010), el AFE y el AFC no son categorías conceptuales distintas, más bien constituyen los dos polos de un continuo exploratorio-confirmatorio, con muchas opciones intermedias, una situación análoga al continuo unidimensionalidad-multidimensionalidad. En realidad, permítasenos la licencia, todas las dicotomías, o casi todas, son meras simplificaciones de un mundo en el que reinan el continuo y la probabilidad.

Otra posibilidad para evaluar la unidimensionalidad es utilizar el modelo bifactorial (MB), en el cual cada indicador, los ítems en el caso de un test, satura en un factor general y en uno, y solo uno, específico (Chen, Hayes, Carver, Laurenceau y Zhang, 2012; Jennrich y Bentler, 2011; Reise, 2012; Wells y Faulkner-Bond, 2016). Incluso cabe una aproximación no paramétrica, como el procedimiento DIMTEST (Hattie, Krakowski, Rogers y Swaminathan, 1996; Stout, 1987; Van Abswoude, Van der Ark y Sijtsma, 2004), o el método DETECT (Gierl, Leighton y Tan, 2006; Stout el al., 1996; Svetina, 2013; Zhang y Stout, 1999).

Al lado de estos indicadores se han propuesto otros muchos complementarios para evaluar la independencia local, que como se ha señalado viene implícita si demostramos previamente la unidimensionalidad. Los interesados pueden acudir al clásico de Yen (1984), o a los trabajos de Chen y Thissen (1997), Ip (2001), Levy, Mislevy y Sinharay (2009) o Liu y Maydeu (2013).

En suma, existen muchas opciones y el software correspondiente para evaluar en qué medida los ítems de un test constituyen una sola dimensión y por tanto son susceptibles de ser analizados mediante un modelo de TRI que asuma la unidimensionalidad. Dado que en la práctica nunca vamos a encontrar una unidimensionalidad pura, es importante combinar los métodos estadísticos utilizados con los criterios y conocimientos derivados del campo sustantivo en el que se trabaja. Los conocimientos teóricos y los métodos estadísticos deben ir de la mano a la hora de la toma de decisiones, tanto en lo relativo a la unidimensionalidad que nos ocupa ahora como en otros aspectos. Los investigadores y los profesionales deben apoyarse en una instrumentación metodológica rigurosa y hacer acopio del mayor número posible de evidencias empíricas, pero al final las decisiones no se delegan en las técnicas utilizadas, hay que tomarlas y correr riesgos, confiando en que al final la ciencia convergerá hacia *la verdad* a base de iteraciones sucesivas llevadas a cabo por investigadores independientes.

4.2. Elección del modelo

Supuesto que los ítems conforman un test unidimensional, el siguiente problema es qué modelo de TRI es más razonable utilizar. En primer lugar, señalar que cualquier elección a priori es lícita para el investigador, y que será el ajuste del modelo a los datos el que decida lo correcto o incorrecto de la elección. No obstante, ciertas características de los ítems pueden proporcionar algunas claves que mejoren la mera elección al azar o capricho. Por ejemplo, es poco razonable intentar ajustar el modelo de un parámetro (Rasch) si se sospechan índices de discriminación no iguales, lo cual puede evaluarse tentativamente escrutando dichos índices en la teoría clásica, o si la probabilidad de acertarlos al azar es considerable. En ambos casos es desaconsejable a priori un modelo de un parámetro, que, como se ha visto, asume un índice de discriminación constante para todos los ítems ($a = K$) y la inexistencia de aciertos al azar ($c = 0$). Asimismo, si $c \neq 0$, el modelo de dos parámetros es igualmente poco plausible a priori. Nótese que cuando los ítems son de elección múltiple, siempre existe cierta probabilidad de aciertos al azar. Por ejemplo, en el caso de cuatro alternativas con solo una correcta, la probabilidad de acierto al azar, aun sin saber nada, es de 1/4 = 0,25, que, aunque no es estrictamente equivalente al parámetro c, no deja de ser una buena aproximación. Teóricamente, el modelo de tres parámetros debería ser preferible a los de uno y dos, ya que constituyen casos particulares de aquel, pero, por contra, el de un parámetro es de cálculo e interpretación sencillos, por lo que en la práctica es el preferido de los usuarios. Incluso es atractivo desde el punto de vista teórico por su parsimonia, al postular que la respuesta de una persona a un ítem solo depende de la competencia de esa persona en la variable medida por el ítem (θ) y de la dificultad del ítem (b). Además, la estima-

ción del parámetro *c* en el modelo de tres parámetros no está lo bien resuelta que sería de desear. En todo caso, no olvidar que los jueces han de ser los datos, eligiéndose aquel modelo que mejor dé cuenta de ellos, preferencias aparte, y, eso sí, en caso de ajustes similares escójase el más sencillo, como mandan los cánones de la parsimonia científica y el sentido común. Elegido el modelo, a continuación han de *estimarse los parámetros* de los ítems y la competencia de cada persona en la variable medida (θ). Finalmente, habrá que comprobar que el modelo con los parámetros así estimados se *ajusta* a los datos empíricos generados por las personas al responder a los ítems.

4.3. Estimación de los parámetros

Seleccionado uno de los modelos, el paso siguiente será estimar los parámetros de cada ítem y el valor de la variable medida (θ) para cada persona a partir de los datos obtenidos al aplicar los ítems a una muestra amplia de personas. Por tanto, la aplicación de los ítems a una muestra representativa precede a la estimación. Con los datos obtenidos, esto es, con las respuestas empíricas de las personas a los ítems, se lleva a cabo la estimación, cuya lógica general consiste en elegir como valores para los parámetros aquellos que maximicen la probabilidad de que ocurran los datos que de hecho se han dado en las respuestas de las personas. Para hacerse una idea elemental y aproximativa de la lógica, imagínese que se dispone de una moneda lastrada de la que se desconoce la probabilidad de obtener cara y cruz (no conocemos esos dos parámetros). Llevamos a cabo el experimento de recogida de datos tirándola al aire 1.000 veces, obteniendo 700 caras y 300 cruces. Pues bien, el valor de la probabilidad de obtener cara que hace más verosímil lo ocurrido es 700/1.000 = 0,70; por tanto, estimamos que el valor (desconocido) de obtener cara con esa moneda es 0,70, según los datos. Este método de estimación se denomina «máxima verosimilitud», en referencia precisamente a que los valores estimados son aquellos que hacen más verosímiles, más plausibles, los datos obtenidos. No obstante, otros métodos son posibles (Lord, 1986; Swaminathan, 1983).

La estimación se va haciendo por aproximaciones sucesivas (iteraciones) y su cálculo es muy laborioso, por lo que es necesaria la ayuda de los ordenadores. El proceso de iteraciones se detiene cuando los valores estimados de los parámetros convergen, esto es, cuando tras una iteración *n* no se producen cambios significativos en los valores estimados. En la actualidad se dispone de distintos programas informáticos para la estimación de los parámetros (véase el cuadro adjunto, en el que se presentan algunos de ellos). La mayoría de los programas ofrecen como salida fundamental los valores estimados de los parámetros de cada ítem y el valor de θ de cada persona, aparte de otros datos importantes como el funcionamiento diferencial de los ítems, la función de información o la curva característica del test, entre otros.

Una forma típica de comprobar lo adecuado de estos programas para estimar los parámetros es mediante simulación por ordenador. Para ello se generan (simulan) las respuestas de las personas a los ítems a partir de parámetros conocidos y luego se les aplica el programa correspondiente, comprobándose en qué grado dicho programa recupera (estima) los parámetros previamente conocidos a partir de los cuales se generaron los datos. Nótese que, mediante esta lógica, lo único que se confirma o falsea es el método de estimación de los parámetros implementado en el programa de ordenador, paso previo a los estudios de validación empírica de los modelos.

Exposiciones detalladas sobre la estimación de los parámetros pueden consultarse en Baker (1987), Birnbaum (1968), Lord (1980) o Swaminathan (1983), siendo especialmente recomendable la de Hambleton y Swaminathan (1985) por su claridad y utilización de ejemplos numéricos sencillos que ayudan a la comprensión. Aquí, basándonos en los trabajos de los autores citados, nos limitaremos a dar una idea introductoria que permita al lector captar la lógica del proceso y acudir seguidamente a las fuentes citadas para un mayor detalle y profundización. Hambleton et al. (1991) y Zhao y Hambleton (2009) ofrecen una excelente clasificación de los programas de ordenador disponibles, analizando sus objetivos, características, pros y contras; para software no comercial, véase Deng (2009). Una buena presentación del software R de libre acceso para psicometría puede verse en el monográfico de la revista *Journal of Statistical Software* (Leew y Mair, 2007), y en español el trabajo de Elosua (2009) constituye una buena introducción a R.

En el cuadro adjunto se reseñan algunos de los programas más relevantes junto con sus autores.

Algunos programas informáticos para estimar los parámetros de los modelos

Programa	Modelos	Fuente
BIGSTEPS	Modelo de Rasch y derivados	Linacre y Wright (1998)
BILOG	1p, 2p, 3p	Mislevy y Bock (1984)
BILOG-MG	1p, 2p, 3p, DIF y equiparación	Zimowski, Muraki, Mislevy y Bock: http://www.ssicentral.com
ConQuest	1p, multicategorial, multidimensional	Adams, Wu y Wilson (2015)
DIMENSION	Multidimensional	Hattie y Krakowski (1994)
flexMIRT	Multidimensional	Cai (2013)
IRTPRO	1p, 2p, 3p, multicategorial	Cai, Thissen y Du Toit (2011)
LOGIST	1p, 2p, 3p	Wingersky (1983). Wingersky et al. (1982)
MICROSCALE	Multicategorial (lp)	Mcdiax Interactivc Technologies (1986)
MIRTE	1p, 2p, 3p	Carlson (1987)
MULTILOG	Multicategoria1	Thissen, Chen y Bock: http://www.ssicentral.com
NOHARM	1p, 2p, 3p, multidimensional	Fraser y McDonald (1988)
PARSCALE	1P, 2P, 3P, multicategorial	Muraki y Bock (1991): http://www.ssicentral.com
PML	1p	Gustafsson (1980)
RASCAL	1p	Assessment System Corporation: http://www.assess.com
RIDA	1p	Glas (1990)
RUMM2030	1p, 2p, 3p, multicategorial	Andrich y Luo: http://www.rummlab.com.au/
Software libre R	1p, 2p, 3p, multicategorial	R Core Team (2014)
WinGen	Generar datos TRI	Han (2007): http://www.hantest.net/wingen
WinSteps	Modelo de Rasch y derivados	Linacre (2015): http://www.winsteps.com
XCALIBRE	1p, 2p, 3p	http://www.assess.com

Estimación condicional y estimación conjunta

Como ya se ha señalado, los valores estimados para los parámetros por el método de máxima verosimilitud serán aquellos que maximicen la probabilidad de ocurrencia de los datos obtenidos al aplicar los ítems a las personas. Si se dispusiera de una función matemática que «representase» dichas respuestas, el problema se reduciría a hallar los máximos de la función y adoptar como valores de los parámetros aquellos que correspondiesen a los puntos donde la función tuviese los máximos, esto es, aquellos que maximizan la función. Veamos cómo se procede para generar esta función.

Supóngase un ítem con una determinada CCI, la cual proporciona la probabilidad que las personas con determinado valor en θ tienen de acertar el ítem. La variable U_i, «respuesta a un ítem», solo tiene dos valores: o se acierta, en cuyo caso $u_i = 1$, o se falla, $u_i = 0$. Precisamente, la CCI informa de la probabilidad de acierto y de fallo para un valor dado de θ:

$$P(U_i = 1|\theta) = P(\theta) \quad \text{y} \quad P(U_i = 0|\theta) = Q(\theta) = 1 - P(\theta)$$

Véase ilustrado en la figura 7.10 para una CCI.

La probabilidad de acertar este ítem para las personas con $\theta = 2$ es 0,88, esto es:

$$P(U_i = 1|\theta = 2) = 0{,}88$$

y la de fallarlo:

$$P(U_i = 0|\theta = 2) = 1 - 0{,}88 = 0{,}12$$

Nótese que esa es la probabilidad; otra cuestión es cuál será de hecho la respuesta de cada persona al ítem para ese nivel de θ: unos lo acertarán y otros lo fallarán; a pesar de la baja probabilidad de que esto ocurra, a la larga se espera que lo acierten el 88 % y lo fallen el 12 %.

Ahora bien, la respuesta a un ítem para un determinado nivel de θ es una prueba de Bernoulli; por tanto:

$$P(U_i|\theta) = P(U_i = 1|\theta)^{U_i} P(U_i = 0|\theta)^{(1-U_i)}$$

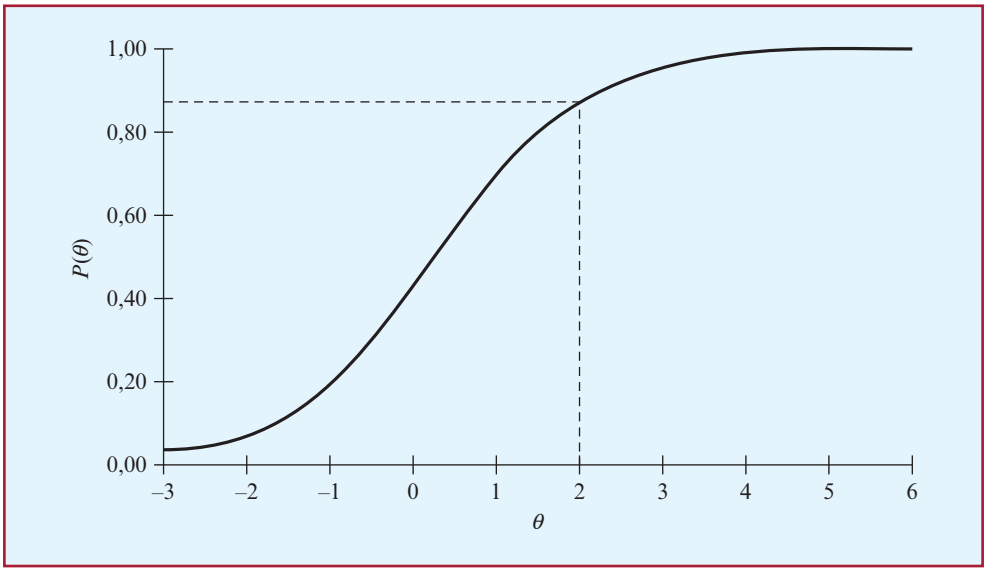


Figura 7.10.—Curva característica de un ítem que ilustra la probabilidad de acertar y fallar el ítem para los distintos valores de θ, y en especial para $\theta = 2$.

y sustituyendo:

$$P(U_i|\theta) = [P(\theta)]^{U_i}[Q(\theta)]^{(1-U_i)} \qquad [7.12]$$

que no es otra cosa que un modo formalizado de expresar lo dicho. En efecto, para el caso anterior en el que $P(\theta) = 0{,}88$ y $Q(\theta) = 0{,}12$, la probabilidad de acertar el ítem viene dada por:

$$P(U_i = 1|\theta = 2) = (0{,}88)^1(0{,}12)^{(1-1)} = 0{,}88$$

y la de fallarlo por:

$$P(U_i = 0|\theta = 2) = (0{,}88)^0(0{,}12)^{(1-0)} = 0{,}12$$

Ahora bien, un test estará compuesto por n ítems, por lo que la probabilidad de que se produzca un determinado patrón de respuestas vendrá dada por el producto de las probabilidades, puesto que se asume su independencia (independencia local). Por ejemplo, en un test con cinco ítems la probabilidad de que al nivel $\theta = 2$ se dé el patrón de respuestas 11010 será:

$$P_1(\theta)P_2(\theta)Q_3(\theta)P_4(\theta)Q_5(\theta)$$

donde los subíndices indican los ítems correspondientes, y $P(\theta)$ y $Q(\theta)$, los valores correspondientes a las CCI de los cinco ítems para ese nivel de θ. Todo lo cual puede expresarse formalmente:

$$L(u_1, u_2, u_3, \ldots, u_n|\theta) = \prod_{i=1}^{n}[P_i(\theta)]^{U_i}[Q_i(\theta)]^{(1-u_i)} \qquad [7.13]$$

expresión que se denomina «función de verosimilitud». Los valores de $P_i(\theta)$ y $Q_i(\theta)$ vienen dados por las CCI y, por tanto, variarán según el modelo de TRI con el que se trabaje.

Si se conocen los parámetros de los ítems, es relativamente fácil estimar el valor de θ de la persona. Este caso se presentaría cuando se dispone de un banco de ítems calibrados (con parámetros ya estimados) y se desea estimar la competencia de una persona. No obstante, no conviene perder de vista que el problema central de la estimación consiste en estimar conjuntamente los parámetros desconocidos de los ítems y la puntuación θ de cada persona a partir de las respuestas a los ítems, siéndonos todo ignoto salvo las respuestas empíricas de las personas. Sí que tiene interés ilustrativo asumir por un momento que se conocen los parámetros de los ítems, pudiendo hallarse bajo tal asunción los

valores de la función de verosimilitud $L(u|\theta)$ para los distintos valores de θ, sustituyéndolos sucesivamente. Se tomará como estimación de la θ de la persona aquel valor de θ para el cual $L(u|\theta)$ tenga un máximo.

En la figura 7.11 se estimaría a la persona considerada una $\theta = -1$, que es donde la función $L(u|\theta)$ tiene su máximo.

En vez de trabajar con $L(u|\theta)$, se suele hacer con su logaritmo, lo cual no altera el valor estimado de θ y tiene la ventaja operativa de convertir el producto en sumas:

$$\ln[L(u|\theta)] = \sum_{i=1}^{n} [u_i \ln P_i(\theta) + (1 - u_i) \ln Q_i(\theta)]$$

Para hallar el máximo se iguala la primera derivada (aquí respecto de θ) a cero: $\delta \ln [L(u|\theta)]/\delta(\theta) = 0$ y el valor de θ será el obtenido al resolver esta ecuación llamada «ecuación de verosimilitud». Esta ecuación es generalmente no lineal y no puede resolverse analíticamente, por lo que han de emplearse procedimientos numéricos, siendo el más habitual el de Newton-Raphson, mediante el cual se llevan a cabo estimaciones sucesivas de θ (iteraciones) hasta que se produce la convergencia, esto es, hasta que el valor estimado en la iteración K no difiere significativamente del obtenido en la iteración $K - 1$.

Para N personas y n ítems, la función de verosimilitud vendrá dada por:

$$L(u|\theta) = \prod_{a=1}^{N} \prod_{i=1}^{n} [P_{ia}(\theta)]^{u_{ia}} [Q_{ia}(\theta)]^{(1-u_{ia})} \qquad [7.14]$$

o en forma logarítmica:

$$\ln [L(u|\theta)] = \sum_{a=1}^{N} \sum_{i=1}^{n} [u_{ia} \ln P_{ia}(\theta) + (1 - u_{ia}) \ln \{1 - P_{ia}(\theta)\}] \qquad [7.15]$$

Los estimadores de máxima verosimilitud de θ para las N personas se obtendrán resolviendo $[\delta \ln L/\delta\theta_a] = 0$ para cada θ_a. Desarrollando, esta expresión puede explicitarse para los modelos logísticos (véase, por ejemplo, Hambleton y Swaminathan, 1985) del siguiente modo:

Modelo logístico de un parámetro

$$\partial \ln L/\partial\theta_a \equiv \sum_{i=1}^{n} Du_{ia} - \sum_{i=1}^{n} DP_{ia}(\theta) = 0 \qquad [7.16]$$

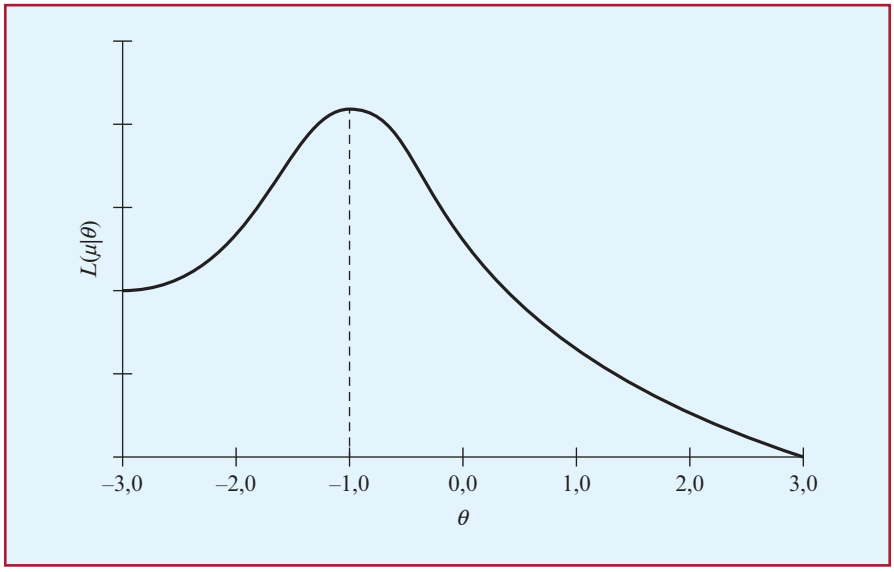


Figura 7.11.—Función de verosimilitud.

Modelo logístico de dos parámetros

$$\partial \ln L/\partial\theta_a \equiv \sum_{i=1}^{n} Da_i u_{ia} - \sum_{i=1}^{n} Da_i P_{ia}(\theta) = 0 \qquad [7.17]$$

Modelo logístico de tres parámetros

$$\partial \ln L/\partial\theta_a \equiv \sum_{i=1}^{n} \frac{[Da_i\{P_{ia}(\theta) - c_i\}(u_{ia})]}{P_{ia}(\theta)(1-c_i)} - \\ - \sum_{i=1}^{n} \frac{[Da_i\{P_{ia}(\theta) - c_i\}P_{ia}(\theta)]}{P_{ia}(\theta)(1-c_i)} \qquad [7.18]$$

La estimación de θ así realizada suele denominarse «estimación condicional de máxima verosimilitud», aludiendo lo de condicional a que la estimación está condicionada al conocimiento previo de los parámetros de los ítems. Ahora bien, en la situación real más frecuente se desconocen los parámetros de los ítems, que habrá que estimar conjuntamente con los valores de θ para cada persona, denominándose «estimación conjunta de máxima verosimilitud». Los parámetros a estimar serán $(n + N)$ en el modelo logístico de un parámetro, es decir, el parámetro b para cada uno de los ítems, lo que hace n parámetros, más un parámetro θ para cada una de las N personas. En el modelo de dos parámetros, los parámetros a estimar serán $(2n + N)$ y en el tres $(3n + N)$. La lógica de la estimación sigue siendo la misma, se elegirán como estimaciones aquellos valores de los parámetros que *conjuntamente* maximicen la función:

$$L(u|\theta, a, b, c) = \prod_{a=1}^{N}\prod_{i=1}^{n}[P_{ia}(\theta)]^{u_{ia}}[Q_{ia}(\theta)]^{(1-u_{ia})} \qquad [7.19]$$

o en forma logarítmica:

$$\ln L(u|\theta, a, b, c) = \sum_{a=1}^{N}\sum_{i=1}^{n}[u_{ia}\ln P_{ia}(\theta) + \\ + (1-u_{ia})\ln Q_{ia}(\theta)] \qquad [7.20]$$

De nuevo, para hallar las estimaciones de θ, a, b y c que maximicen la función L hay que resolver:

$$[\partial \ln L/\partial P_K] = 0$$

donde P es el vector de parámetros a estimar, $P' = [\theta, a, b, c]$ y K el número de estos. Hambleton y Swaminathan (1985, p. 132) ofrecen los valores de esta expresión para el modelo logístico de tres parámetros:

$$\partial \ln L/\partial a_i = \frac{D}{1-c_i}\sum_{a=1}^{N}\frac{(\theta_a - b_i)[P_{ia}(\theta) - c_i][u_{ia} - P_{ia}(\theta)]}{P_{ia}(\theta)}$$

$$\partial \ln L/\partial b_i = \frac{-Da_i}{1-c_i}\sum_{a=1}^{N}\frac{[P_{ia}(\theta) - c_i][u_{ia} - P_{ia}(\theta)]}{P_{ia}(\theta)}$$

$$\partial \ln L/\partial c_i = \frac{1}{1-c_i}\sum_{a=1}^{N}\frac{[u_{ia} - P_{ia}(\theta)]}{P_{ia}(\theta)}$$

$$\partial \ln L/\partial \theta_a = D\sum_{i=1}^{n}\frac{a_i[P_{ia}(\theta) - c_i][u_{ia} - P_{ia}(\theta)]}{(1-c_i)P_{ia}(\theta)}$$

En el caso de estimación condicionada, las ecuaciones se resolvían una a una, independientemente, pero aquí ha de hacerse conjuntamente, por lo que se requiere la utilización de un tratamiento multivariado del procedimiento de Newton-Raphson. La estimación se lleva a cabo en dos pasos: primero se estiman las puntuaciones θ de las personas, según lo expuesto en el caso condicionado, tomando los parámetros de los ítems como conocidos y asignándoles un valor inicial. Las θ estimadas se asumen conocidas y en un segundo paso se estiman los parámetros de los ítems. Ambos se repiten hasta obtener una convergencia conjunta y los valores con los que, finalmente, se logra la convergencia se toman como los estimadores de máxima verosimilitud.

Cabe señalar, finalmente, que, además del tipo de estimaciones aquí reseñadas de máxima verosimilitud, cada vez cobran más fuerza las aproximaciones bayesianas, las cuales al incorporar distribuciones a priori permiten una mejor estimación de los parámetros (Gifford y Swaminathan, 1990; Lord, 1986; Mislevy, 1986; Swaminathan, 1983; Swaminathan y Gifford, 1982, 1985, 1986; Tsutakawa y Lin, 1986).

4.4. Ajuste del modelo

Una vez estimados los parámetros del modelo, hay que comprobar en qué grado los resultados pronosticados con esos valores coinciden con los obte-

nidos de hecho; en otras palabras, hay que comprobar el ajuste del modelo a los datos. Si tal ajuste se produce, ello quiere decir que los valores de $P(\theta)$ pronosticados por el modelo no difieren estadísticamente de los obtenidos empíricamente, es decir, la proporción de personas que de hecho aciertan el ítem.

Existen varios procedimientos estadísticos para la comprobación del ajuste, si bien ninguno de ellos es totalmente satisfactorio. Buenos tratamientos pueden consultarse en Orlando y Thissen (2000), Stone y Zhang (2003) o Haberman, Sinharay y Chon (2013). Para una aproximación no paramétrica, véanse Douglas y Cohen (2001), Wells y Bolt (2008), Liang y Wells (2009), Liang, Wells y Hambleton (2014) o Wells, Rios y Faulkner-Bond (2016). En español, López Pina e Hidalgo (1996).

Aquí se ilustrará con fines didácticos la lógica del ajuste mediante el uso de chi-cuadrado, el análisis de los residuos y la comparación *de las distribuciones de las puntuaciones.*

Chi-cuadrado

La lógica general de las técnicas basadas en chi-cuadrado consiste en comparar los valores pronosticados por el modelo con los obtenidos empíricamente. Para ello se divide el rango de la variable medida θ en varias categorías y se comparan los valores pronosticados y empíricos para cada categoría. En el caso de ajuste perfecto, ambos valores coincidirán en todas las categorías; a medida que aumentan las diferencias, el ajuste es peor. Precisamente lo que nos indicará chi-cuadrado es si esas diferencias son estadísticamente significativas. Aquí se ilustrarán el estadístico Q_2 propuesto por Wright y Panchapakesan (1969) para el modelo logístico de un parámetro y el estadístico Q_1 de Yen (1981) aplicable a cualquiera de los tres modelos logísticos.

Wright y Panchapakesan (1969) propusieron un estadístico sencillo para comprobar el ajuste de los modelos a los datos, cuya distribución se aproxima a la de Q_2:

$$Q_2 = \sum_{j=1}^{k} \frac{n_j[P(\theta_j) - P_e(\theta_j)]^2}{[P(\theta_j)][1 - P(\theta_j)]} \qquad [7.21]$$

k: Número de categorías en las que se divide θ.

n_j: Número de personas dentro de cada categoría.

$P(\theta_j)$: Valor de la CCI dado por la fórmula del modelo con los parámetros estimados, para la categoría j.

$P_e(\theta_j)$: Proporción de personas que, de hecho (empíricamente), superan el ítem para una categoría determinada j.

Q_2: Se distribuye según χ^2 con $k - 1$ grados de libertad.

Ejemplo

En una muestra de 1.000 personas se estimó mediante un programa que el modelo que mejor se ajustaba a los datos obtenidos al aplicar un test de 20 ítems era el logístico de un parámetro. En concreto, para el ítem 10 el programa asignó $b = 2$. El número de personas que acertaron el ítem 10 para las categorías en las que se dividió θ aparecen en la tabla adjunta. A la vista de los resultados, ¿puede afirmarse al nivel de confianza del 99% que el modelo se ajusta a los datos para el caso del ítem 10?

$\hat{\theta}$	n_j	$P_e(\theta_j)$
4-5	70	0,97
3-4	90	0,95
2-3	200	0,70
1-2	300	0,35
0-1	340	0,10
	1.000	

Para poder aplicar la fórmula propuesta hay que obtener previamente los valores de $P(\theta_j)$ dados por el modelo. Recuérdese que el modelo logístico de un parámetro viene dado por:

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$

Sustituyendo los valores correspondientes al ítem 10 ($b = 2$):

$$P(\theta) = \frac{e^{D(\theta - 2)}}{1 + e^{D(\theta - 2)}}$$

como $D = 1{,}7$ y $e = 2{,}72$:

$$P(\theta) = \frac{(2{,}72)^{(1{,}7)(\theta - 2)}}{1 + (2{,}72)^{(1{,}7)(\theta - 2)}}$$

Sustituyendo θ por los valores centrales de las categorías en las que se dividió $\hat{\theta}$, se obtienen los correspondientes $P(\theta_j)$:

$\hat{\theta}$	$P(\theta_j)$
4,5	0,99
3,5	0,92
2,5	0,70
1,5	0,30
0,5	0,07

Aplicando la fórmula:

$$Q_2 = \frac{340(0{,}07 - 0{,}10)^2}{(0{,}07)(1 - 0{,}07)} + \frac{300(0{,}30 - 0{,}35)^2}{(0{,}30)(1 - 0{,}30)} + \frac{200(0{,}70 - 0{,}70)^2}{(0{,}70)(1 - 0{,}70)} + \frac{90(0{,}92 - 0{,}95)^2}{(0{,}92)(1 - 0{,}92)} + \frac{70(0{,}99 - 0{,}97)^2}{(0{,}99)(1 - 0{,}99)} =$$

$$= 4{,}70 + 3{,}57 + 0{,}00 + 1{,}10 + 2{,}83 = 12{,}2$$

Ahora bien, en las tablas $\chi^2_{0{,}99}$, con $k - 1 = 5 - 1 = 4$ grados de libertad es igual a 13,28; por tanto, al nivel de confianza del 99% se admite (no se puede rechazar) que el modelo así estimado se ajusta a los datos para el ítem 10.

En la figura 7.12 aparecen representados los valores $P(\theta_j)$ pronosticados por el modelo y los obtenidos empíricamente $P_e(\theta_j)$.

Nótese que propiamente los valores de θ generados por el programa utilizado son estimaciones de θ, de ahí el «sombrero» ($\hat{\theta}$), los verdaderos valores de θ no se conocen. Sobre la métrica de θ se hablará más adelante; de momento se ha prescindido de los valores negativos por sencillez.

CASO DE N ÍTEMS

Si se desea someter a prueba el ajuste del modelo no para un ítem determinado, como en el caso anterior, sino para los n ítems que componen el test conjuntamente, y obtener así una idea global del funcionamiento del modelo, los autores citados (Wright y Panchepakesan, 1969) proponen un estadístico generalización del anterior:

$$Q_{2T} = \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{n_j [P(\theta_{ji}) - P_e(\theta_{ji})]^2}{P(\theta_{ji})[1 - P(\theta_{ji})]} \qquad [7.22]$$

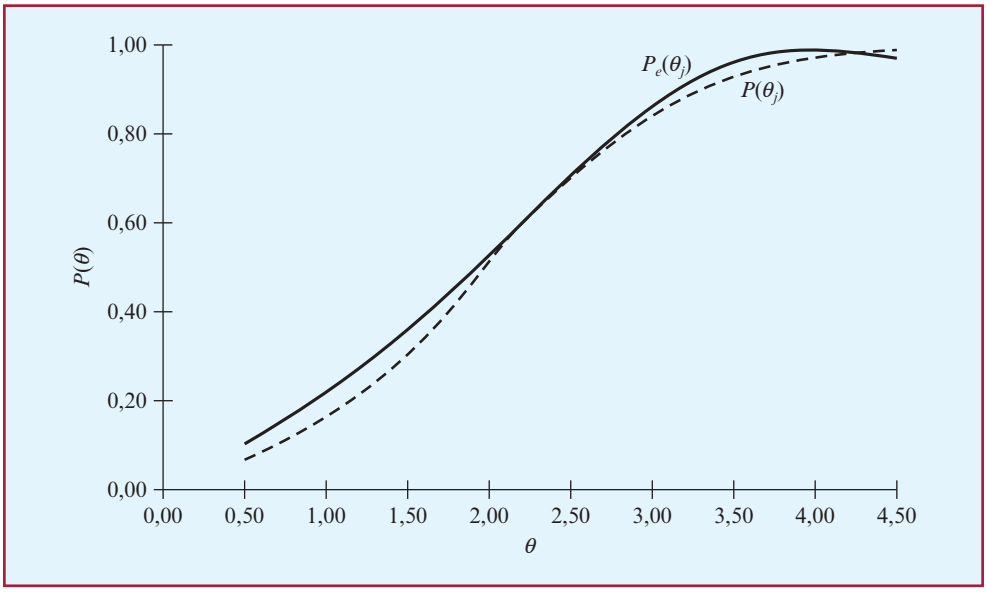


Figura 7.12.—Valores pronosticados por el modelo y valores empíricos obtenidos.

donde todos los términos son idénticos a los ya citados, n es el número de ítems y Q_{2T} se distribuye según χ^2 con $(k-1)(n-1)$ grados de libertad. Se trata, por tanto, de realizar el mismo proceso anterior para cada ítem, sumar los resultados y contrastarlos con el χ^2 crítico de las tablas. Afortunadamente, este y otros métodos de ajuste son proporcionados por los programas informáticos, y la única finalidad de nuestro ejemplo es eso, ejemplificar la lógica del procedimiento.

Cabe preguntarse cuál es el número más adecuado de categorías en el que ha de dividirse θ. No hay una respuesta definitiva, siendo frecuente utilizar entre 10 y 15. En el modelo de Rasch, dado que a cada puntuación empírica se le estima una θ, si se dispone de un número elevado de personas, es razonable hacer una categoría para cada puntuación. Nótese, por contra, que en los modelos de dos y tres parámetros una misma puntuación en el test no necesariamente recibe la misma estimación θ, que depende del patrón de respuestas a los ítems, no solo del número de aciertos, lo cual, aunque no supone un inconveniente teórico, sí lo es práctico, pues la lógica más corriente del usuario y del que responde a un test es que a puntuaciones iguales (número de ítems acertados) corresponden competencias iguales.

Yen (1981) propuso un estadístico similar al anterior aplicable a cualquiera de los tres modelos logísticos:

$$Q_1 = \sum_{j=1}^{k} \frac{n_j[P(\theta_j) - P_e(\theta_j)]^2}{P(\theta_j)[1 - P(\theta_j)]} \qquad [7.23]$$

donde todos los términos ya han sido definidos para el estadístico anterior y Q_1 se distribuye según χ^2 con $k - p$ grados de libertad, siendo k el número de categorías en las que se dividió θ y p el número de parámetros del modelo de TRI utilizado.

Aplicado a los datos del ejemplo anterior, se obtiene igual que entonces un valor de $Q_1 = 12{,}2$. En este caso, los grados de libertad serían también $5 - 1 = 4$, pues se trata del modelo logístico de un parámetro; luego $p = 1$.

Si bien el ejemplo anterior trata de ilustrar la lógica del ajuste, el estadístico Q_1 tiene varias limitaciones (Wells et al., 2016), por lo que en la actualidad existen mejores opciones, como las citadas en las referencias recomendadas más arriba.

Análisis de los residuos

Otro modo muy parejo a los anteriores de acercarse a la computación del ajuste del modelo a los datos es el análisis de los residuos. Como antes, se divide θ en varias categorías o niveles y se calcula para cada una de ellas el residuo estandarizado (RE):

$$\text{RE} = \frac{P(\theta_j) - P_e(\theta_j)}{\sqrt{P(\theta_j)Q(\theta_j)/n_j}} \qquad [7.24]$$

donde:

- n_j: Número de personas dentro de la categoría j.
- $P(\theta_j)$: Valor de la CCI para el nivel θ_j.
- $P_e(\theta_j)$: Proporción empírica de personas dentro de una categoría dada j que superan el ítem.
- $Q(\theta_j)$: $1 - P(\theta_j)$.

A medida que los residuos se alejan de cero en valor absoluto, peor será el ajuste del modelo. Una inspección del tamaño de los residuos para las distintas categorías en las que se dividió θ puede dar una idea descriptiva de las zonas de mayor desajuste del ítem. Por ejemplo, véase en la figura 7.13 el tipo de residuos que daría un modelo mal ajustado tal como el que se representa.

En los primeros niveles, los residuos son negativos, $P(\theta_j) < P_e(\theta_j)$, para en las últimas invertirse la relación. Más que pruebas estadísticas rigurosas, como sería de desear, es frecuente que los investigadores establezcan una banda de valores admisibles para los residuos, por ejemplo, entre –2 y +2, u otros valores arbitrarios pero interpretables asumiendo la distribución normal de los residuos. El programa informático ResidPlots (Liang, Han y Hambleton, 2009) permite obtener datos detallados sobre los residuos, y es compatible con la mayoría del software de TRI, como PARSCALE, BILOG-MG o MULTILOG.

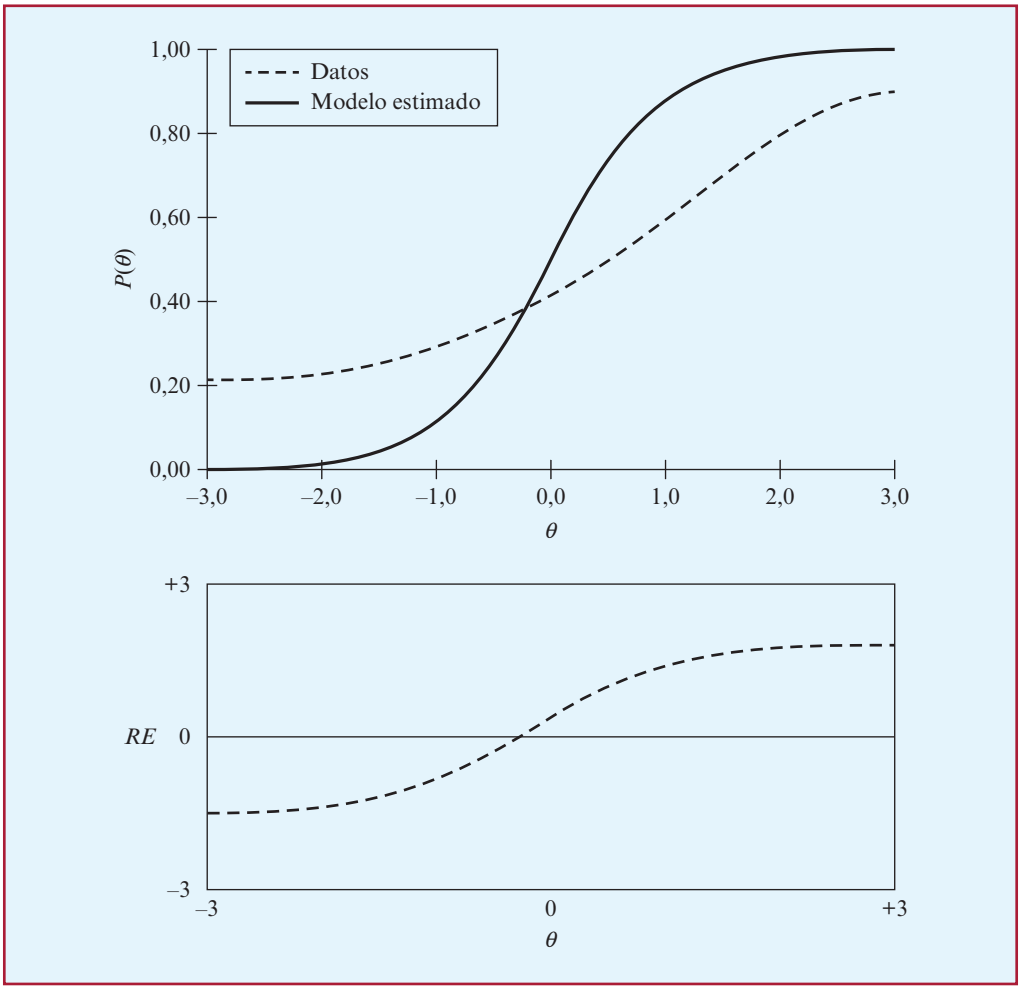


Figura 7.13.—Discrepancia entre el modelo estimado y los datos empíricos.

Comparación de las distribuciones

Un tercer tipo de índice de ajuste de los modelos consiste en comparar, mediante las técnicas estadísticas al uso, por ejemplo, χ^2, las distribuciones de las puntuaciones empíricas del test y la distribución teórica generada por la curva característica del test, concepto que se verá más adelante.

4.5. Invarianza de los parámetros

Si se superan los pasos anteriores con éxito —como se ha visto, cada uno de ellos envuelve una problemática peculiar—, finalmente se dispondrá de un *test calibrado,* es decir, de un conjunto de ítems con sus parámetros adecuadamente estimados (listos para aplicar a nuevas personas) y de una muestra de N personas, los utilizados en el estudio, de los que se conoce su nivel estimado en la variable medida θ. No hay que confundir el valor de θ para una persona con su puntuación en el test. Como se verá más adelante, la curva característica del test permitirá establecer la relación entre θ y las puntuaciones de las personas en el test.

Ahora bien, cabe preguntarse por la conexión entre un modelo así establecido y los objetivos de invarianza prometidos por la TRI, a saber:

a) Estimar las puntuaciones de las personas sin que importe el instrumento utilizado.
b) Estimar los parámetros de los ítems independientemente de la muestra empleada.

Es algo que conviene entender cabalmente, pues constituye el meollo de la TRI. Si el modelo se ajusta estrictamente a los datos, los dos objetivos se cumplen (véase figura 7.14).

En la figura 7.14 aparecen las CCI de tres ítems correspondientes a un modelo logístico de tres parámetros. Los valores de los parámetros a, b y c de los ítems son, como se puede observar, diferentes. Sobre el eje de abscisas se han representado las distribuciones de θ para dos muestras de personas evaluadas, N_1 y N_2. La estimación del valor de θ, sea θ_j, para una persona o clase de personas determinada; nótese que no depende de que utilicemos un tipo de ítem u otro, lo único que variará será la $P(\theta)$ según la forma de la CCI dada por el valor de sus parámetros. Así, por ejemplo, si a una persona (o grupo de personas) con $\theta = \theta_j$ se le aplican tres test, el primero con 100 ítems del tipo 1, el segundo con 100 ítems del tipo 2 y el tercero con 100 del tipo 3, todos ellos, obviamente, midiendo la misma variable θ, acertaría, respectivamente, 25, 35 y 65 ítems, ya que, según el gráfico, $P_1(\theta_j) = 0{,}25$, $P_2(\theta_j) = 0{,}35$ y $P_3(\theta_j) = 0{,}65$, pero en los tres casos se le asignará idéntica θ_j.

En segundo lugar, los parámetros de los ítems no dependen del tipo de muestra, los valores de $P(\theta)$ no están en función de la distribución de θ para las personas. Véase en la figura 7.14, por ejemplo, cómo θ_j genera idénticos valores $P(\theta)$ tanto se considere en la muestra N_1 como en la N_2, con distribuciones muy diferentes. No obstante, como en cualquier otro caso de estimación estadística, cuanto mayor sea la muestra de personas y mejor cubran el rango de valores de θ, más precisas serán las estimaciones de los parámetros.

Comprobación de la invarianza

Para comprobar la invarianza de la estimación de la θ de cada persona para distintos test que miden la misma variable, se aplican dos (o más) test compuestos por distintos ítems a la misma muestra de personas y luego se ve en qué grado ambas estimaciones coinciden. Dicha coincidencia puede indagarse representando gráficamente las θ obtenidas en un test frente a las obtenidas en el otro: cuanto

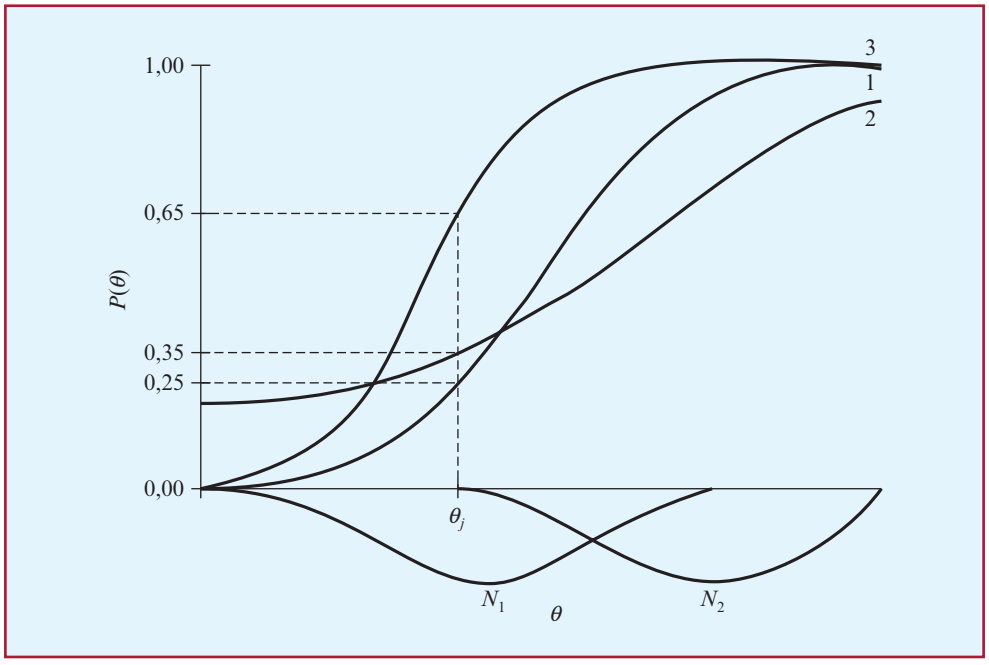


Figura 7.14.—Curvas características de tres ítems estimadas a partir de dos muestras con distribuciones diferentes en θ.

más se acerquen a una recta, más fina será la invarianza. Una indicación numérica del ajuste puede hallarse mediante la correlación de Pearson entre las estimaciones.

En la figura 7.15 aparecen representados los valores de θ para 100 personas en dos test distintos, con una correlación entre ambas estimaciones de 0,96, que puede considerarse un buen ajuste. Ha de notarse que, aunque las estimaciones de θ con test distintos fuesen exactamente las mismas para cada persona, ello no quiere decir que el programa informático empleado dé el mismo valor numérico en cada caso; de ahí la necesidad del gráfico y la correlación. La razón es muy simple: no existe una métrica única para θ; por tanto, el programa establece para cada análisis una métrica en función de los parámetros de los ítems utilizados y, en consecuencia, el valor de θ de una persona depende de esa métrica, así que, aunque la invarianza sea perfecta, la salida del ordenador no tiene por qué ser la misma para una persona en ambas ocasiones, pero sí tiene que haber una relación lineal perfecta entre las estimaciones. De cómo transformar dos estimaciones en métricas distintas a una misma métrica se tratará en el capítulo dedicado a la equiparación de las puntuaciones.

Invarianza de los parámetros de los ítems. Análogamente, si se utilizan diferentes muestras para estimar los parámetros de *n* ítems, el modelo postula la invarianza de estos. *Mutatis mutandis,* la comprobación empírica es similar a la anterior, ahora son los parámetros de los ítems los que han de compararse en vez de θ. En la figura 7.16 aparecen representadas las estimaciones del parámetro *b* (índice de dificultad) de 50 ítems en dos muestras de personas. La invarianza es notable: $r_{12} = 0{,}98$. Algunos ejemplos numéricos pueden consultarse en Hambleton y Swaminathan (1985) y Lord (1980).

El *parámetro c* (aciertos al azar) no viene afectado por la elección del origen de la escala y de sus unidades, luego su estimación ha de ser idéntica para ambas (o más) muestras (Lord, 1980).

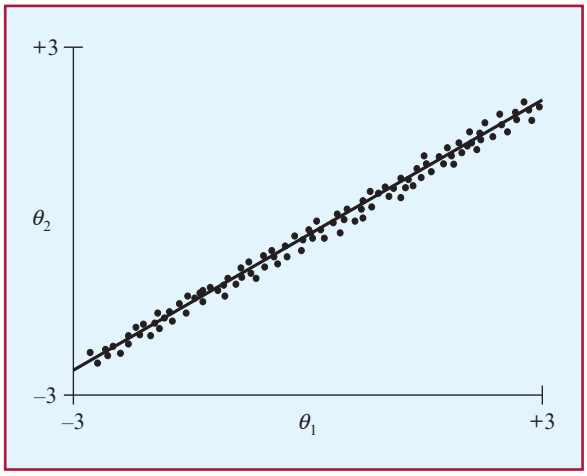


Figura 7.15.—Representación de los valores de θ estimados por dos tests distintos.

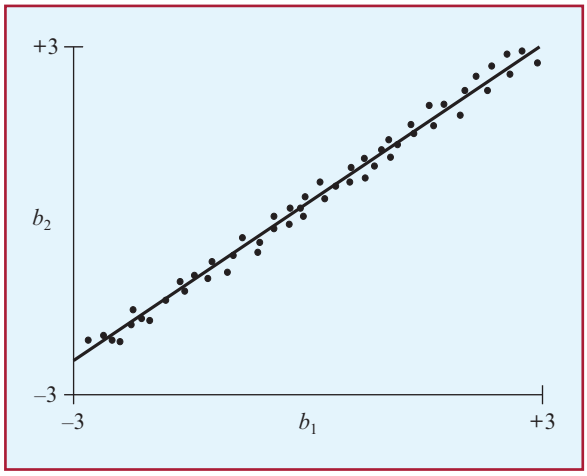


Figura 7.16.—Valores de *h* estimados en dos muestras distintas de personas.

EJERCICIOS

1. Se aplicó un test de cinco ítems a una muestra de 100 personas. Para cada ítem se calcularon los índices clásicos de dificultad (ID), discriminación (r_b) y la probabilidad de acertar el ítem al azar (P_a), cuyos valores aparecen en la tabla adjunta.

Ítems	ID	r_b	P_a
1	0,50	0,80	0,00
2	0,10	0,20	0,20
3	0,60	0,50	0,25
4	0,70	0,35	0,00
5	0,90	0,10	0,15

1. Identifique el ítem más fácil y el más difícil.
2. ¿Cuál es el ítem más discriminativo? ¿Y el menos?
3. ¿Cuáles de los ítems de este test es más probable que no sean de elección múltiple?
4. ¿Qué modelo de TRI elegiría para utilizar con este test?

2. Se aplicó un test de 40 ítems a una muestra de 200 personas, cuyas puntuaciones θ en el test se estimaron utilizando el modelo de Rasch. Las puntuaciones θ se dividieron en seis intervalos para estudiar el ajuste del modelo a los datos. En la tabla adjunta aparece el número de personas comprendidas en cada intervalo (n_j), así como las que de ellas acertaron el ítem séptimo del test (f_a).

θ	n_j	f_a
–3 - –2	20	2
–2 - –1	30	9
–1 - 0	40	20
0 - 1	60	51
1 - 2	40	36
2 - 3	10	10

1. El parámetro b para el ítem séptimo cuyos datos aparecen en la tabla fue –0,5. Calcule el estadístico chi-cuadrado de Wright y Panchapakesan y diga si el modelo se ajusta a los datos para el ítem séptimo. Nivel de confianza, 95%.
2. Calcule los residuos estandarizados (RE) para el ítem séptimo (datos de la tabla). Si se adopta como criterio de un buen ajuste que ninguno de los residuos supere el valor absoluto de 1, ¿puede afirmarse que el modelo se ajusta a los datos?
3. A la vista de los residuos estandarizados del apartado anterior, ¿en qué zona de θ se produce un mejor ajuste del modelo a los datos para este ítem?

3. A continuación aparecen los valores de los parámetros b estimados para seis ítems en dos muestras distintas de personas.

	Muestra 1	Muestra 2
Ítems	b_1	b_2
1	–1,50	–1,30
2	0,25	0,20
3	1,20	1,10
4	1,50	1,40
5	2,00	2,02
6	2,50	0,00

1. Tomando como criterio la correlación entre las estimaciones halladas en ambas muestras, la invarianza podría considerarse: muy pobre, moderada, excelente.
2. Elimine el ítem cuya invarianza parece más defectuosa y recalcule la correlación entre las estimaciones de b para el resto de los ítems. ¿Cómo calificaría la invarianza una vez descartado el citado ítem?
3. Transforme los valores de b_1 y b_2 a una nueva escala E', según la siguiente expresión:

$$E' = 100(b) + 500$$

4. ¿Qué cambios se han producido en la invarianza de b tras la transformación del apartado anterior?

4. En la tabla adjunta aparecen las respuestas de 60 personas a un ítem de un test de inteligencia. Las 60 personas se han organizado en seis intervalos

θ	Personas									
–3 - –2	0	1	0	0	0	0	0	0	1	0
–2 - –1	1	0	0	0	1	0	0	1	0	0
–1 - 0	1	0	0	0	1	1	0	1	1	0
0 - 1	1	1	1	0	0	1	1	0	1	0
1 - 2	1	1	1	1	0	1	1	1	0	0
2 - 3	1	1	1	1	1	1	1	0	0	1

según sus puntuaciones en la variable medida θ. Los parámetros estimados para el ítem de la tabla mediante el modelo logístico de tres parámetros fueron: $a = 0{,}6$, $b = 0{,}5$ y $c = 0{,}2$.

1. Calcule los valores $P(\theta_j)$ pronosticados por el modelo para el punto medio de cada una de las categorías de θ.
2. Al nivel de confianza del 95%, ¿puede afirmarse que el modelo se ajusta a los datos para el ítem de la tabla? (utilice chi-cuadrado).
3. Si se toma como criterio de ajuste que ninguno de los residuos estandarizados del ítem supere un valor absoluto de 1,96, ¿puede afirmarse que el ítem de la tabla se ajusta a los datos?

SOLUCIONES

1.1. Fácil: 5; Difícil: 2.
2. Más discriminativo: 1; menos discriminativo: 5.
3. 1 y 4.
4. Modelo logístico de 3 parámetros.

2.1. $\chi^2 = 15{,}46 > 11{,}07$ (tablas): el modelo no se ajusta a los datos.
2. Ajuste deficiente, tres de los cinco RE tienen un valor absoluto superior a 1.
3. Mejor ajuste en la zona central de θ, entre –1 y +1, donde los RE = 0.

3.1. $r(b_1, b_2) = 0{,}71$. Invarianza moderada.
2. $r(b_1, b_2) = 0{,}99$. Invarianza excelente.
3. $b_1' = 350, 525, 620, 650, 700, 750$. $b_2' = 370, 520, 610, 640, 702, 500$.
4. Ninguno, la transformación lineal no afecta a la invarianza.

4.1. 0,23; 0,29; 0,41; 0,60; 0,79; 0,91.
2. Se ajusta: $Q_1 = 2{,}35 < 7{,}81$.
3. Se ajusta: todos los residuos estandarizados son menores que 1,96 (0,21, 0,07, 0,64, 0,00, 0,64, 1,10).

5. MÉTRICA DE *THETA* (θ)

Pocos tópicos de la TRI han sido tan malinterpretados como el de la métrica de θ. La frase habitual es que en los modelos de TRI la métrica de θ es arbitraria, lo cual es cierto, siempre y cuando se defina bien lo que se entiende por arbitrario. El lector que no se haya saltado los apartados anteriores, si hubiese alguno, ya habrá captado la idea general respecto a esta arbitrariedad, consistente en que el modelo establecido no predetermina el origen ni las unidades de θ, esto es, sigue siendo válido, *siguen obteniéndose las mismas* $P(\theta)$, si se utiliza otro origen y otras unidades para medir θ, siempre y cuando, claro está, se hagan las oportunas correcciones en los parámetros de los ítems. Veamos primero un ejemplo numérico y luego se concretarán las transformaciones de θ admisibles para los modelos logísticos.

Ejemplo

Supóngase un modelo logístico de dos parámetros, formulado en cierta métrica θ, con $a = 1{,}5$ y $b = 2$; luego vendrá dado por:

$$P(\theta) = \frac{e^{D(1{,}5)(\theta - 2)}}{1 + e^{D(1{,}5)(\theta - 2)}}$$

y como $e = 2{,}72$ y $D = 1{,}7$:

$$P(\theta) = \frac{(2{,}72)^{(1{,}7)(1{,}5)(\theta - 2)}}{1 + (2{,}72)^{(1{,}7)(1{,}5)(\theta - 2)}} = \frac{(2{,}72)^{(2{,}55)(\theta - 2)}}{1 + (2{,}72)^{(2{,}55)(\theta - 2)}}$$

Para un valor de $\theta = 2$, sustituyendo se obtendrá un valor de

$$P(\theta) = \frac{(2{,}72)^{(2{,}55)(2 - 2)}}{1 + (2{,}72)^{(2{,}55)(2 - 2)}} = 0{,}5$$

Ahora supóngase que los valores de la escala θ se transforman, por ejemplo, multiplicándolos todos por 3 y sumándoles 1, es decir, el valor anterior de $\theta = 2$ quedará transformado en:

$$\theta' = 3(2) + 1 = 7$$

Veamos que si se sustituye ahora en el modelo el nuevo valor $\theta' = 7$, se obtendrá el mismo resultado que al sustituir el original de $\theta = 2$. Ahora bien, si se ha transformado la escala de θ, antes de hacer los cálculos habrá que transformar también los valores originales de $b = 2$ y $a = 1{,}5$ en los correspondientes en la nueva escala:

$$b' = 3(2) + 1 = 7$$

$$a' = \frac{1{,}5}{3} = 0{,}5$$

Por tanto:

$$P(\theta) = \frac{(2{,}72)^{(1{,}7)(0{,}5)(7-7)}}{1 + (2{,}72)^{(1{,}7)(0{,}5)(7-7)}} = 0{,}5$$

Es decir, $P(\theta) = P(\theta') = 0{,}5$. Ese es el sentido de la arbitrariedad de la métrica de θ: podemos elegir las puntuaciones linealmente transformadas que más nos interesen sin alterar los valores de $P(\theta)$ proporcionados por los modelos. Veamos a continuación las transformaciones admisibles de θ para los modelos logísticos de uno, dos y tres parámetros.

5.1. Transformaciones admisibles de θ

Modelo logístico de un parámetro

Si se suma (o resta) una constante k a todos los valores de θ y se hace lo mismo con el índice de dificultad b, no se altera el valor de $P(\theta)$.

Sea $\theta' = \theta + k$, $b' = b + k$:

$$P(\theta') = \frac{e^{D(\theta' - b')}}{1 + e^{D(\theta' - b')}} = \frac{e^{D[(\theta+k)-(b+k)]}}{1 + e^{D[(\theta+k)-(b+k)]}} =$$

$$= \frac{e^{D(\theta - b)}}{1 + e^{D(\theta - b)}} = P(\theta)$$

En suma, $P(\theta') = P(\theta)$.

Si se añade un factor multiplicativo M, además del aditivo anterior, para mantener la invarianza hay que modificar el modelo, introduciendo una nueva constante $1/M$ que multiplica a D. Por tanto, tal como se expresa aquí, el modelo no es invariante a una transformación lineal multiplicativa (aunque es posible cierta «acomodación» para esta transformación manipulando el valor medio del índice de discriminación asignado a los ítems).

Nótese que lo que implica esta invarianza de $P(\theta)$ respecto a la adición de una constante es que una persona no tiene una θ determinada ni un ítem una b específica, siempre es posible transformar dichos valores en otros sin alterar el modelo. Dada esta inespecificidad, los programas informáticos suelen elegir como métrica de θ aquella con media cero y desviación típica 1. El usuario es luego libre de transformar estas puntuaciones en otras que considere más oportunas.

Modelo logístico de dos parámetros

$P(\theta)$ resulta invariante a cualquier transformación lineal de θ siempre que esta se aplique también al índice de dificultad b y que el índice de discriminación a se divida por el factor multiplicativo. Es decir:

$$\theta' = M(\theta) + k$$
$$b' = M(b) + k$$

$$a' = \frac{a}{M}$$

Efectivamente:

$$P(\theta') = \frac{e^{Da'(\theta' - b')}}{1 + e^{Da'(\theta' - b')}} =$$

$$= \frac{e^{D(a/M)[(M\theta+k)-(Mb+k)]}}{1 + e^{D(a/M)[(M\theta+k)-(Mb+k)]}} =$$

$$= \frac{e^{D(a/M)M(\theta - b)}}{1 + e^{D(a/M)M(\theta - b)}} = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}} = P(\theta)$$

(Véase el ejemplo numérico desarrollado al comienzo del apartado.)

Modelo logístico de tres parámetros

Análogo al de dos parámetros, pero además con $c' = c$.

$$\theta' = M(\theta) + k$$
$$b' = M(b) + k$$
$$a' = \frac{a}{M}$$
$$c' = c$$
$$P(\theta') = \frac{c' + (1 - c')e^{Da'(\theta' - b')}}{1 + e^{Da'(\theta' - b')}} =$$
$$= \frac{c + (1 - c)e^{D(a/M)[(M\theta + k) - (Mb + k)]}}{1 + e^{D(a/M)[(M\theta + k) - (Mb + k)]}} =$$
$$= \frac{c + (1 - c)e^{D(a/M)M(\theta - b)}}{1 + e^{D(a/M)M(\theta - b)}} =$$
$$= \frac{c + (1 - c)e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}} = P(\theta)$$

Esta indeterminación de la escala de θ, como ya se ha señalado, obliga en cada situación de calibración de un test a elegir una, con las consecuencias citadas sobre la métrica de los parámetros de los ítems, siendo lo más corriente que la métrica elegida por los programas ubique la media en cero y la desviación típica en 1. Si se desea evitar los valores negativos y decimales, se puede llevar a cabo cualquier otra transformación lineal admisible.

5.2. Transformaciones de $P(\theta)$: *logits*

Además de las transformaciones lineales admisibles citadas de la escala de θ, es frecuente usar en vez de los valores directos de $P(\theta)$ alguna modificación *no lineal* de ellos. La razón es tratar de captar más cabalmente la significación del modelo, aunque si no se entienden bien tales modificaciones, en vez de claridad, añaden oscuridad. Una de las transformaciones más utilizadas son los así llamados *logits*. Veamos a qué se denomina logit.

Modelo logístico de un parámetro

Como es bien sabido, el modelo viene dado por

$$P(\theta) = \frac{e^{D(\theta - b)}}{1 + e^{D(\theta - b)}}$$

Por tanto:

$$Q(\theta) = 1 - P(\theta) = 1 - \frac{e^{D(\theta - b)}}{1 + e^{D(\theta - b)}} =$$
$$= \frac{1 + e^{D(\theta - b)} - e^{D(\theta - b)}}{1 + e^{D(\theta - b)}} = \frac{1}{1 + e^{D(\theta - b)}}$$

Dividiendo en el modelo original ambos miembros por $Q(\theta)$:

$$\frac{P(\theta)}{Q(\theta)} = \frac{[e^{D(\theta - b)}]/[1 + e^{D(\theta - b)}]}{1/[1 + e^{D(\theta - b)}]} = e^{D(\theta - b)}$$

Si se hace $D = 1$ o, al modo de Rasch-Wright, su valor se incluye en θ y b, es decir, se hace la multiplicación, entonces:

$$\frac{P(\theta)}{Q(\theta)} = e^{(\theta - b)}$$

Tomando logaritmos neperianos:

$$\ln \frac{P(\theta)}{Q(\theta)} = (\theta - b) \ln (e)$$

como $\ln (e) = 1$:

$$\ln \frac{P(\theta)}{Q(\theta)} = (\theta - b) \qquad [7.25]$$

Se denomina logit a la unidad de la escala $\ln [P(\theta)/Q(\theta)]$, es decir, al logaritmo neperiano del cociente entre la probabilidad de pasar el ítem según el modelo y la probabilidad de fallarlo. Un logit $(\theta - b) = 1$, o, lo que es lo mismo, $\ln [P(\theta)/Q(\theta)] = 1$, indicará que $P(\theta)/Q(\theta) = 2{,}72$, ya que $\ln (2{,}72) = 1$. En otras palabras, un valor de 1 en la escala logit equivale a 2,72 en la escala $P(\theta)/Q(\theta)$.

La escala logit proporciona una cierta idea a la hora de comparar personas entre sí. Sea, por ejemplo, una persona con $\theta = \theta_1$ y otro con $\theta = \theta_2$ y un ítem j con un índice de dificultad $b = b_j$:

$$\ln \frac{P(\theta_1)}{Q(\theta_1)} = (\theta_1 - b_j)$$

$$\ln \frac{P(\theta_2)}{Q(\theta_2)} = (\theta_2 - b_j)$$

Restando miembro a miembro:

$$\ln \frac{P(\theta_1)}{Q(\theta_1)} - \ln \frac{P(\theta_2)}{Q(\theta_2)} = (\theta_1 - b_j) - (\theta_2 - b_j)$$

Ahora bien, la diferencia de logaritmos es igual al logaritmo del cociente:

$$\ln \frac{P(\theta_1)/Q(\theta_1)}{P(\theta_2)/Q(\theta_2)} = (\theta_1 - \theta_2) \qquad [7.26]$$

Las diferencias entre las dos personas se transforman directamente en diferentes probabilidades de acertar el ítem. Así, por ejemplo, si $(\theta_1 - \theta_2) = 0$, ello significa que:

$$\frac{P(\theta_1)/Q(\theta_1)}{P(\theta_2)/Q(\theta_2)} = 1$$

Obviamente, si no hay diferencias entre la competencia de las personas, las probabilidades de éxito de ambos serán iguales y, en consecuencia, su cociente 1. Cuando $(\theta_1 - \theta_2) = 1$, entonces el cociente será 2,72; para una diferencia de –0,223 el cociente vendrá dado por 0,8, y para una diferencia de 1,25 valdrá 3,5, etc. Todo lo cual informa adecuadamente de la probabilidad que tienen comparativamente ambas personas de superar el ítem.

Modelo logístico de dos parámetros

$$P(\theta) = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

$$Q(\theta) = 1 - P(\theta) = 1 - \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}} =$$

$$= \frac{1 + e^{Da(\theta - b)} - e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}} = \frac{1}{1 + e^{Da(\theta - b)}}$$

Dividiendo $P(\theta)$ entre $Q(\theta)$:

$$\frac{P(\theta)}{Q(\theta)} = \frac{[e^{Da(\theta - b)}]/[1 + e^{Da(\theta - b)}]}{1/[1 + e^{Da(\theta - b)}]} = e^{Da(\theta - b)}$$

Tomando logaritmos neperianos:

$$\ln \frac{P(\theta)}{Q(\theta)} = Da(\theta - b) \ln (e) = Da(\theta - b) \qquad [7.27]$$

Por tanto, para este modelo el logit incluye el valor del índice de discriminación a.

Si se desea comparar a dos personas con $\theta = \theta_1$ y $\theta = \theta_2$, respectivamente, en un ítem b_j, los logits vendrán dados por:

$$\ln \frac{P(\theta_1)}{Q(\theta_1)} = Da_j(\theta_1 - b_j)$$

$$\ln \frac{P(\theta_2)}{Q(\theta_2)} = Da_j(\theta_2 - b_j)$$

Restando miembro a miembro:

$$\ln \frac{P(\theta_1)}{Q(\theta_1)} - \ln \frac{P(\theta_2)}{Q(\theta_2)} = Da_j(\theta_1 - b_j) - Da_j(\theta_2 - b_j)$$

Simplificando:

$$\ln \frac{P(\theta_1)/Q(\theta_1)}{P(\theta_2)/Q(\theta_2)} = Da_j(\theta_1 - \theta_2) \qquad [7.28]$$

Modelo logístico de tres parámetros

$$P(\theta) = c + (1 - c) \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

Pasando c al primer miembro:

$$P(\theta) - c = (1 - c) \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

$$Q(\theta) = 1 - P(\theta) = 1 - \left[c + (1-c)\frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}} \right] =$$

$$= (1-c) - (1-c)\frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}} =$$

$$= (1-c)\left(1 - \frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}\right) =$$

$$= (1-c)\frac{1+e^{Da(\theta-b)} - e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}} = \frac{1-c}{1+e^{Da(\theta-b)}}$$

En los modelos de uno y dos parámetros se dividía $P(\theta)$ entre $Q(\theta)$, pero aquí se divide $P(\theta)$ corregido por los aciertos al azar c; es decir, se divide $[P(\theta) - c]$ entre $Q(\theta)$; por tanto:

$$\frac{P(\theta)-c}{Q(\theta)} = \frac{(1-c)[e^{Da(\theta-b)}]/[1+e^{Da(\theta-b)}]}{(1-c)/[1+e^{Da(\theta-b)}]} = e^{Da(\theta-b)}$$

Tomando logaritmos neperianos:

$$\ln\frac{P(\theta)-c}{Q(\theta)} = Da(\theta-b)\ln(e) = Da(\theta-b) \qquad [7.29]$$

Aplicado a la comparación de dos personas, análogamente al caso de dos parámetros:

$$\ln\frac{[P(\theta_1)-c]/Q(\theta_1)}{[P(\theta_2)-c]/Q(\theta_2)} = Da(\theta_1-\theta_2) \qquad [7.30]$$

Las derivaciones anteriores pueden llevarse a cabo con distintas bases logarítmicas, así como extenderse a la comparación entre ítems, además de entre personas (véase, por ejemplo, Hambleton y Swaminathan, 1985).

5.3. Otras transformaciones

Otra escala de interés práctico es la de unidades denominadas *wits,* en la cual θ se transforma en θ' según la expresión:

$$\theta' = (10)\log_3 e^{\theta} + 100 \qquad [7.31]$$

La escala de Wits goza de algunas propiedades prácticas deseables.

Finalmente, una transformación interesante es la propuesta por Lord (1980), en la que

$$\theta' = Ke^{k\theta}$$
$$b' = Ke^{kb}$$
$$a' = \frac{Da}{K}$$

donde K y k son constantes positivas. En la nueva escala (véase Lord, 1980, p. 84), para un modelo logístico de tres parámetros se da una sencilla relación:

$$\frac{P(\theta')-c}{Q(\theta')} = (\theta'/b')^{a'} \qquad [7.32]$$

La transformación realizada convierte $P(\theta)$ en otra función $P(\theta')$ que ya no es la logística, pero, como señala el propio Lord, la relación anterior es tan simple y directa que tal vez la escala θ' sea preferible a θ para la medición. A modo de ejercicio, trate el lector de derivar la relación anterior, sustituyendo en el modelo logístico de tres parámetros los valores correspondientes a la transformación y dividendo luego $[P(\theta') - c]$ entre $Q(\theta')$.

EJERCICIOS

1. Utilizando el modelo logístico de tres parámetros, se estimaron los parámetros a, b y c de todos los ítems de un test de inteligencia espacial y las puntuaciones θ de las personas. Las puntuaciones θ obtenidas por las cinco personas de la muestra utilizada fueron las siguientes: –2,4, –1,2, 0,0, 1,6, 2,0.

1. Para un determinado ítem se estimó un índice de dificultad b = 1,5, un índice de discriminación a = 0,8 y un valor de c = 0,20. Estimar la probabilidad que tienen de superar el ítem las personas con una puntuación θ = 1.

2. Transforme las puntuaciones θ de las personas a otra escala de acuerdo con la siguiente expresión: $\theta' = 15(\theta) + 100$.
3. ¿Cuáles son los valores de los parámetros *a, b* y *c* del ítem en la nueva escala?
4. Una persona que hubiese obtenido en la nueva escala una puntuación θ' de 115 puntos, ¿qué probabilidad tiene de superar el ítem?
5. ¿Qué puntuación θ obtuvo en la escala original una persona que en la escala transformada θ' tiene 145 puntos?

2. Tras aplicar un test de 10 ítems a una muestra de 500 personas, se estimaron los parámetros de los ítems mediante el modelo logístico de dos parámetros.

1. Los parámetros de uno de los ítems fueron $a = 0{,}8$ y $b = 1{,}4$. Una persona obtuvo en el test una puntuación $\theta = 1{,}9$; ¿qué puntuación le corresponderá en la escala logit?
2. Si en el ítem anterior el logit correspondiente a una de las personas fuese 1,5, ¿qué puntuación θ habría obtenido en el test esa persona?
3. La diferencia entre las puntuaciones θ de dos personas en el test fue de 3 puntos. ¿Cuál es su diferencia expresada en la escala logit?
4. En otro de los ítems del test cuyos parámetros *a* y *b* desconocemos, a una persona le corresponde una puntuación de 2 en la escala logit. ¿Cuál es la probabilidad de que esa persona supere el ítem? ¿Cuál la de que lo falle?
5. Si el ítem del apartado anterior tuviese una dificultad $b = 1$ y una discriminación $a = 0{,}5$, ¿qué puntuación θ habría obtenido la persona cuyo logit era 2?

SOLUCIONES

1.1. 0,47.
2. 64, 82, 100, 124, 130.
3. $a' = 0{,}05$, $b' = 122{,}5$, $c' = 0{,}2$.
4. 0,47.
5. 3.

2.1. 0,68.
2. 2,5.
3. 4,08.
4. 0,88, 0,12.
5. 3,35.

6. CURVA CARACTERÍSTICA DEL TEST

6.1. Definición

Análogamente al concepto de CCI, pieza central de los modelos de TRI, puede hablarse de curva característica del test (CCT). Aunque su papel en la TRI no es comparable con el de la CCI, tiene gran interés como *puente* entre algunos aspectos de la teoría clásica de los test y la TRI, como ayuda para *interpretar los resultados,* o en la *equiparación* de las puntuaciones, por citar lo más sobresaliente.

La curva característica del test no es otra cosa que la suma de las curvas características de los ítems que componen el test: si para cada nivel de θ se suman los valores de $P(\theta)$ de cada ítem para ese nivel se obtiene la CCT, lo cual puede expresarse así:

$$\text{CCT} = \sum_{i=1}^{n} P_i(\theta) \qquad [7.33]$$

siendo *n* el número de ítems.

Nótese que las sumas han de hacerse para cada nivel de θ y, dado que θ es continua, propiamente habrá que utilizar el cálculo infinitesimal, aunque en la práctica es habitual dividir θ en cortos intervalos sumando las $P(\theta)$ de los ítems para cada intervalo. Se ilustra a continuación con un ejemplo para valores discretos de θ.

Ejemplo

Sea un test compuesto de cuatro ítems cuyos parámetros en un modelo logístico de dos parámetros estimados con el programa BILOG resultaron ser: $a_1 = 1$, $a_2 = 1{,}5$, $a_3 = 2$, $a_4 = 2{,}5$; $b_1 = 0{,}75$, $b_2 = 1$, $b_3 = 2$, $b_4 = 3$. Hallar la CCT. [La suma de la $P(\theta)$ se haría para los valores de θ: –3, –2, –1, 0, 1, 2, 3.]

Sustituyendo los valores a, b y θ en el modelo se obtienen los valores de $P(\theta)$ para los cuatro ítems y se suman para obtener la CCT.

Ítems	$P(\theta)$				$\sum_{i=1}^{n} P_i(\theta)$
	Ítem 4	Ítem 3	Ítem 2	Ítem 1	CCT
–3	0,0000	0,0000	0,0000	0,0017	0,0017
–2	0,0000	0,0000	0,0004	0,0091	0,0095
–1	0,0000	0,0000	0,0059	0,0481	0,0540
0	0,0000	0,0010	0,0719	0,2177	0,2906
1	0,0001	0,0310	0,5000	0,6049	1,1369
2	0,0138	0,5000	0,9280	0,8938	2,3356
3	0,5000	0,9680	0,9940	0,9788	3,4408

En la figura 7.17 aparecen representadas las cuatro CCI, y en la 7.18, la curva característica del test.

6.2. Puntuaciones verdaderas en el test

Como ya se ha señalado, el interés de la TRI no se centra, como ocurría en la teoría clásica de los test, en la estimación de las puntuaciones verdaderas de las personas en el test, sino en la estimación más general de θ, de la que un test particular sería un indicador. No obstante, es ilustrativo entender lo que bajo el prisma de la TRI sería la puntuación verdadera de una persona en el test.

La *puntuación verdadera* en el test de una persona o personas a las que se ha estimado mediante un modelo de TRI una determinada puntuación $\theta = \theta_j$

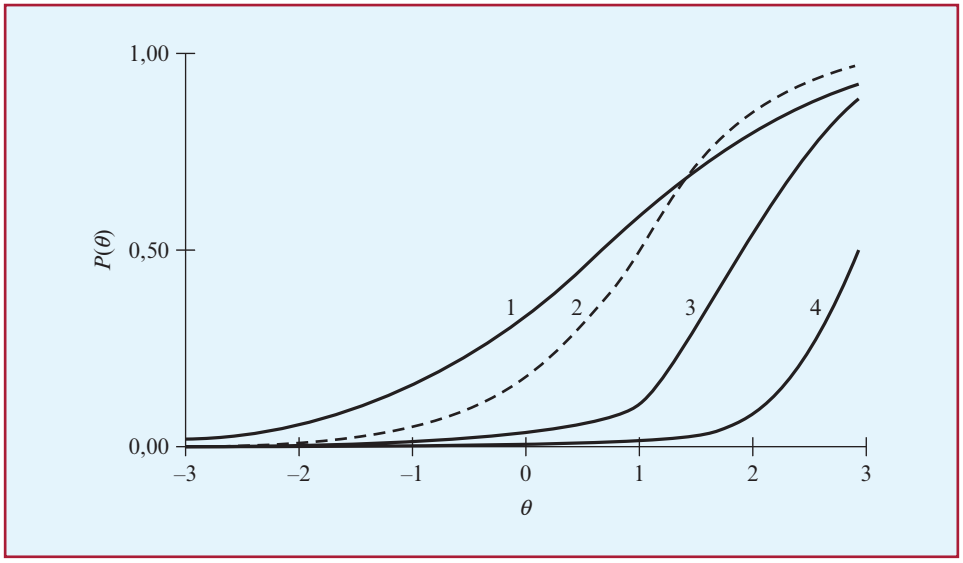


Figura 7.17.—Curvas características de los ítems del ejemplo.

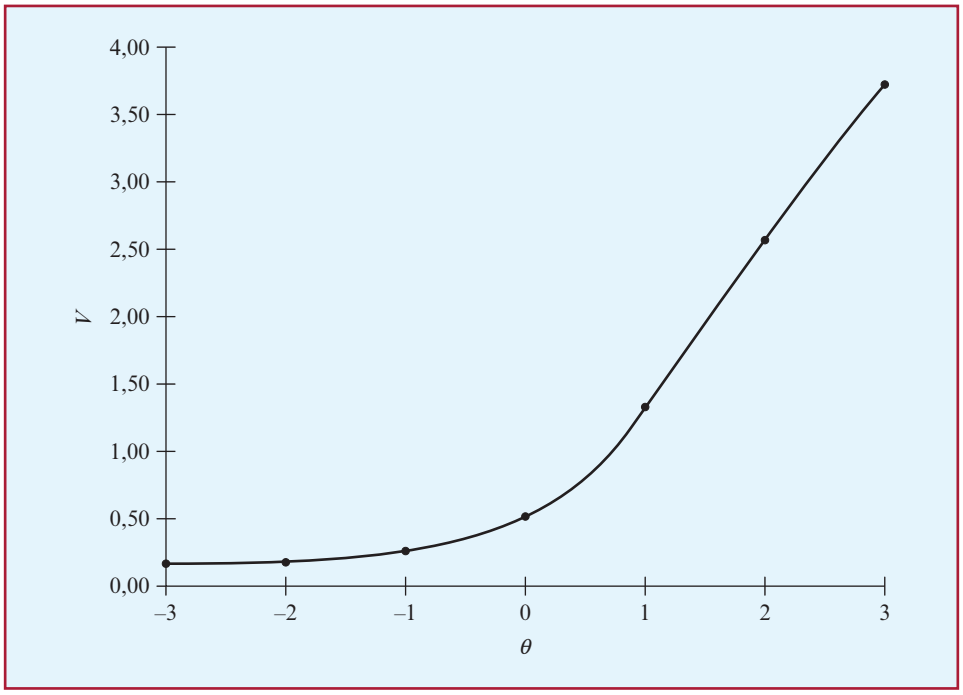


Figura 7.18.—Curva característica del test formado por los cuatro ítems del ejemplo.

viene estimada por la suma de las probabilidades $P(\theta_j)$ dadas por las curvas características de los ítems que componen el test, para el valor de θ_j:

$$V_j = \sum_{i=1}^{n} P_i(\theta_j) \qquad [7.34]$$

donde n es el número de ítems, y $P_i(\theta_j)$, el valor correspondiente a cada CCI para $\theta = \theta_j$. Adviértase que este valor no es otra cosa que el valor generado por la curva característica del test para $\theta = \theta_j$. Por ejemplo, para el test del ejemplo anterior compuesto por cuatro ítems, la puntuación verdadera en el test para las personas a las que se estimó una $\theta = 2$ vendría dada por:

$$V = 0{,}0138 + 0{,}5000 + 0{,}9280 + 0{,}8938 = 2{,}3356$$

Las puntuaciones verdaderas en el test así estimadas pueden ser muy útiles de cara a la interpretación de los resultados, pues *vienen expresadas en la misma escala que las empíricas,* mientras que los valores de θ constituyen otra escala, precisamente una escala que la CCT transforma en puntuaciones verdaderas. Puede afirmarse, por tanto, como señala Lord (1980), que *las puntuaciones verdaderas (V) y las puntuaciones (θ) son la misma cosa pero expresada en diferente escala.* Ahora bien, la gran ventaja a favor de θ es que si el modelo TRI funciona, la puntuación θ estimada a un persona no depende del test utilizado, mientras que V sí. Véase esto ilustrado en la figura 7.19, en la que aparecen las CCT de dos test. Nótese cómo para cada test varía la puntuación verdadera estimada a las mismas personas con $\theta = \theta_j$.

Error típico de medida

Las puntuaciones verdaderas (V), como ocurre en la teoría clásica, y en general en cualquier proceso de estimación, no coincidirán siempre con las empíricas (X), definiéndose el *error de medida* como la diferencia entre ambas ($e = X - V$), y el *error típico de medida* (S_e), como la desviación típica de dichas diferencias.

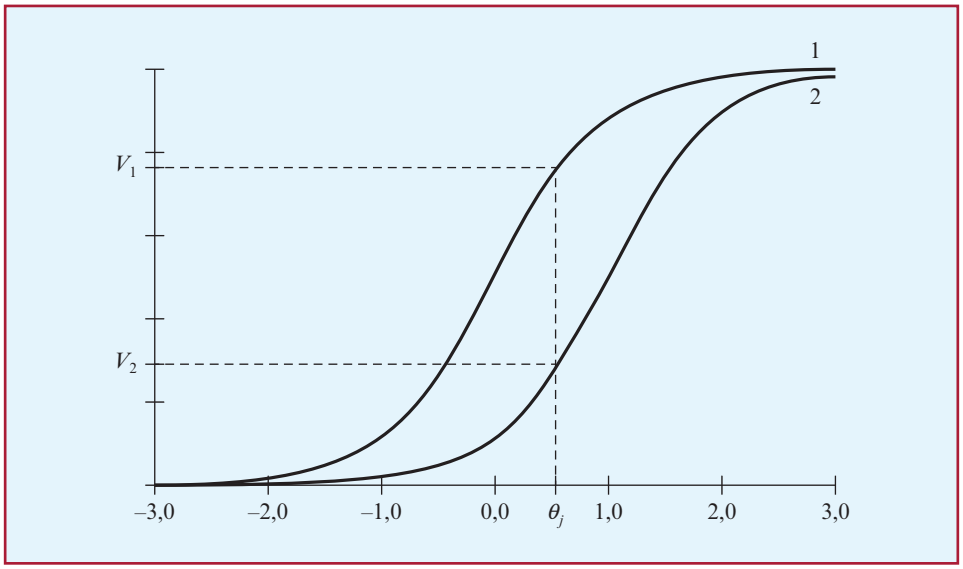


Figura 7.19.—Estimación de la misma θ_j mediante dos curvas características de test diferentes.

Para un cierto nivel de $\theta = \theta_j$ el valor del error típico de medida al cuadrado viene dado por

$$S_e^2 = \sum_{i=1}^{n} P_i[(\theta_j)Q_i(\theta_j)] \qquad [7.35]$$

donde:

n: Número de ítems del test.
$P_i(\theta_j)$: Valor de las CCI para $\theta = \theta_j$, es decir, para el nivel de θ para el cual se desea calcular S_e.
$Q_i(\theta_j) = [1 - P_i(\theta_j)]$.

Este error típico tiene una característica notable respecto al de la teoría clásica: *su valor no es el mismo para todas las personas,* está en función del valor de θ, lo cual quiere decir que la precisión con la que miden los test no es uniforme a lo largo de la escala, va a depender del nivel de las personas en la variable medida.

La obtención de la fórmula del error típico es inmediata. Para cada nivel θ_j de θ, la varianza de los errores de medida de un ítem es la misma que la varianza de las puntuaciones empíricas, pues $\theta = x - v$, con v constante para ese nivel dado θ_j. Para ítems dicotómicos la varianza de cada uno al nivel θ_j vendrá dada por $P(\theta_j)Q(\theta_j)$; luego la varianza de los n ítems que componen el test, para el nivel θ_j, y asumiendo el principio de independencia local, vendrá dada por la suma de las varianzas de los ítems:

$$\sum_{i=1}^{n} P_i(\theta_j)Q_i(\theta_j)$$

que es la fórmula propuesta. Si todos los ítems tuviesen la misma $P(\theta)$, la varianza total de los errores de medida para cada nivel de θ sería la de distribución binomial: $nP(\theta)Q(\theta)$.

6.3. Curva característica de la persona

Análogamente a los conceptos de curva característica del ítem y curva característica del test, puede hablarse de curva característica de la persona (CCP). Se obtiene empíricamente representando en abscisas la dificultad de los ítems (parámetro b), y en ordenadas, la proporción de ítems acertados por la persona en cada categoría (véase la figura 7.20).

En la figura 7.20 aparecen las CCP empíricas de dos personas para el mismo test. Para obtenerlas, en primer lugar, se han de estimar los parámetros b, que se agrupan por categorías según sus valores, en

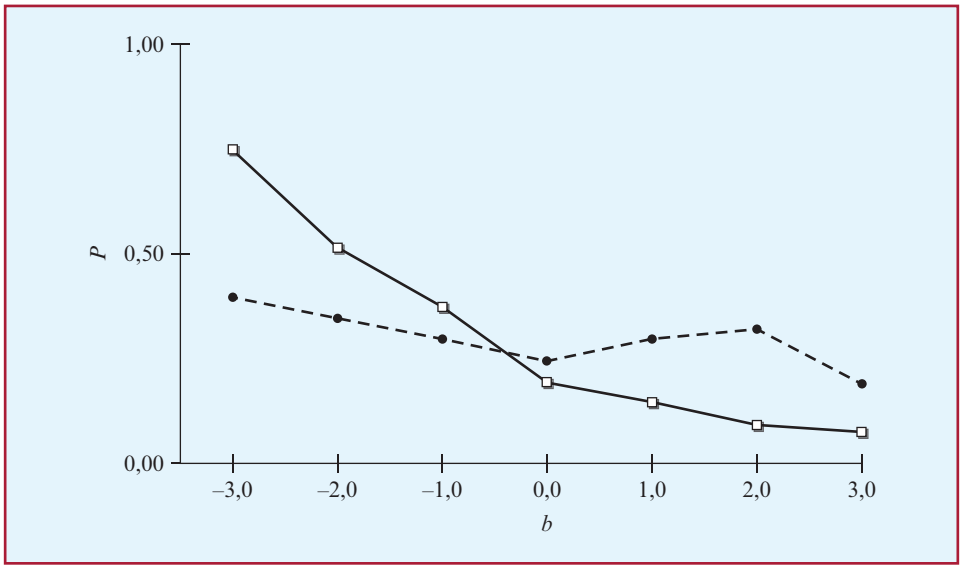


Figura 7.20.—Curvas características empíricas de dos personas.

el ejemplo de –3,5 a + 3,5 (siete categorías). Los valores representados en ordenadas (P) son la proporción de ítems acertados por la persona dentro de cada categoría. Uno de los usos más prometedores de la CCP (Trabin y Weiss, 1983) es su comparación con la curva teórica esperada, lo que permite analizar las discrepancias entre el comportamiento real de una persona ante los ítems y el esperado teóricamente a partir de los parámetros estimados por el modelo de TRI. Para obtener la CCP teórica o esperada se procede como en el caso de la empírica, pero en el eje de ordenadas, en vez de la proporción de aciertos empíricos, se representan los que debería obtener según las correspondientes CCI. En concreto, en ordenadas se ubicará la media de los valores de θ, el estimado para la persona en cuestión, esto es, en ordenadas se representa

$$\sum_{i=1}^{j} \frac{P_i(\theta)}{j}$$

donde j es el número de ítems de cada categoría y $P_i(\theta)$ son los valores correspondientes de las CCI para los ítems en cada categoría al nivel θ de la persona. Por tanto, las CCP esperadas dependen del nivel de la persona en la variable medida, θ, y de las CCI.

EJERCICIOS

1. Un test está formado por cinco ítems cuyos parámetros de dificultad y discriminación estimados mediante el modelo logístico de dos parámetros fueron los siguientes:

Ítems	*a*	*b*
1	1,00	1,25
2	1,50	1,50
3	2,00	2,50
4	2,50	3,50
5	0,50	–1,50

1. Obtenga la curva característica del test, utilizando los siguientes siete valores de θ: –2,5, –1,5, –0,5, 0,5, 1,5, 2,5, 3,5.

2. Según la curva característica del test calculada en el apartado anterior, ¿qué puntuación se estima que obtendrán en el test dos personas a las que se estimó unas puntuaciones θ de 2,5 y 0,5?
3. ¿Cuál es la probabilidad de que una persona a la que se estimó una puntuación $\theta = 3{,}5$ supere los cinco ítems del test?
4. Dos personas obtuvieron, respectivamente, puntuaciones θ de 2,5 y –0,5. ¿Cuál se estima que será la diferencia entre sus puntuaciones en el test?
5. Del test de cinco ítems se suprimieron los dos ítems más difíciles, de modo que quedó reducido a solo tres. ¿Cuál es la probabilidad de que una persona con una puntuación $\theta = 1{,}5$ supere los tres ítems del nuevo test? ¿Qué puntuación se estima que obtendrá esa persona en el test?

2. Un test consta de 40 ítems cuyos índices de dificultad se distribuyen según aparece en la tabla adjunta, donde n_j indica el número de ítems en cada uno de los seis intervalos en los que se clasificaron los ítems atendiendo a su dificultad. Los números de las columnas (*A*, *B* y *C*) reflejan el número de ítems de cada categoría que superaron las personas *A*, *B* y *C* a las que se aplicó el test.

b	n_j	*A*	*B*	*C*
–3 - –2	4	4	3	3
–2 - –1	7	5	4	3
–1 - 0	9	6	5	4
0 - 1	10	6	6	6
1 - 2	6	3	2	4
2 - 3	4	1	0	3

1. Trace las curvas características de la persona para las personas *A*, *B* y *C*.
2. ¿Cuál de las tres curvas características es más atípica? ¿Por qué?
3. A continuación se ofrecen los valores correspondientes a cada categoría provenientes de sumar las curvas características de los ítems. Para obtener estos valores se han sumado las $P_i(\theta)$ de los ítems de cada categoría *j*: $\Sigma P_i(\theta)$.

b	$\Sigma P_i(\theta)$
–3 - –2	3,6
–2 - –1	4,9
–1 - 0	4,5
0 - 1	3,5
1 - 2	1,5
2 - 3	0,4

Obtenga los valores correspondientes a la curva característica esperada y represéntela gráficamente.

4. Compare cada una de las curvas (*A*, *B*, *C*) con la esperada del apartado anterior. Para ello reste en cada una de las categorías el valor esperado del obtenido y pondérelo por el número de ítems de la categoría. (Prescinda del signo de las diferencias, utilice valores absolutos.) Sume posteriormente los valores así obtenidos en cada categoría para obtener un índice global de desajuste de la curva de cada persona respecto de la esperada.

   a) ¿Cuál es el valor del índice así hallado para cada persona?
   b) ¿A qué persona pertenece la curva con peor ajuste a la esperada?

SOLUCIONES

1.1.

θ	CCT
–2,5	0,2198
–1,5	0,5095
–0,5	0,7545
0,5	1,1361
1,5	2,0644
2,5	3,3033
3,5	4,4267

2. 3,3033; 1,1361.

3. 0,46.
4. 2,5488.
5. 0,28; 2,03.

2.1. P_A: 1,00, 0,71, 0,67, 0,60, 0,50, 0,25.
P_B: 0,75, 0,57, 0,55, 0,60, 0,33, 0,00.
P_C: 0,50, 0,43, 0,44, 0,60, 0,67, 0,75.
2. C; acierta en mayor proporción los ítems difíciles.
3. CC esperada: 0,90, 0,70, 0,50, 0,35, 0,25, 0,10.
4. *a*) *A*: 6,60; *B*: 5,34; *C*: 11,65; *b*) *C*.

7. FUNCIÓN DE INFORMACIÓN

7.1. Error típico de estimación de θ

Como ya se ha señalado reiteradamente, los modelos de TRI permiten estimar el valor de θ para todas las personas por el procedimiento de máxima verosimilitud. Una propiedad interesante de estos estimadores (Birnbaum, 1968; Lord, 1980; Hambleton y Swaminathan, 1985) es que se distribuyen asintóticamente normales con media θ y varianza:

$$\text{var}(\hat{\theta}|\theta) = \frac{1}{\sum_{i=1}^{n} \{[P_i'(\theta)]^2 / P_i(\theta)Q_i(\theta)\}} \qquad [7.36]$$

donde:

n: Número de ítems del test.
$P_i(\theta)$: Valores de las CCI de los ítems.
$Q_i(\theta)$: Igual a $1 - P_i(\theta)$.
$P_i'(\theta)$: Derivada de $P_i(\theta)$.

Por tanto, el *error típico* vendrá dado por la raíz cuadrada (desviación típica) de la expresión anterior y permitirá establecer *intervalos confidenciales* en torno a $\hat{\theta}$ para tratar de «apresar» el valor paramétrico («verdadero») de θ.

EJEMPLO

Un test de 20 ítems se aplicó a una muestra de 100 personas. Sabiendo que el error típico de estimación fue 0,20 para un valor estimado de $\hat{\theta} = 2$, calcular al nivel de confianza del 95% entre qué valores se estima que se encontrará θ.

1. NC 95%: $Z_c = \pm 1{,}96$.
2. Error máximo: $(Z_c)(S_e) = (1{,}96)(0{,}20) = = 0{,}392$.
3. $(\hat{\theta} - \text{Error máximo}) \leqslant \theta \leqslant (\hat{\theta} + \text{Error máximo})$
$(2 - 0{,}392) \leqslant \theta \leqslant (2 + 0{,}392)$
$1{,}608 \leqslant \theta \leqslant 2{,}392$.

Luego el valor paramétrico de θ se encontrará entre 1,608 y 2,392, al nivel de confianza del 95%. Nótese que este error típico *no* es el mismo para todas las personas, depende de su nivel en θ.

También dependerá del modelo, pues $[P_i'(\theta)]^2$, $P_i(\theta)$ y $Q_i(\theta)$ varían para un mismo valor de θ según se adopte un modelo u otro. [Más adelante se explicita el valor de $P_i'(\theta)$ para los modelos logísticos de uno, dos y tres parámetros.]

En suma, la TRI permite estimar el valor de θ y proporciona una medida de la precisión de las estimaciones dada por el error típico de estimación. No obstante, para expresar esta información acerca de la precisión de las estimaciones de θ se va a utilizar más que el error típico la *función de información,* que no es otra cosa que el citado error típico expresado de otro modo, como ahora veremos.

7.2. Función de información del test

Birnbaum (1968) define la función de información del test, $I(\theta)$, como el denominador de la fórmula anterior de la varianza del estimador de máxima verosimilitud de θ:

$$I(\theta) = \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad [7.37]$$

donde:

n: Número de ítems.
$P_i'(\theta)$: Derivada de $P_i(\theta)$.

$P_i(\theta)$: Valores de las CCI de los ítems.
$Q_i(\theta)$: Igual a $1 - P_i(\theta)$.

Por tanto, la función de información de un test para un determinado valor de θ es la inversa de la varianza de sus errores de estimación para ese valor:

$$S_e^2 = \text{var}\,(\hat{\theta}|\theta) = \frac{1}{I(\theta)}$$

o, lo que es lo mismo:

$$\text{var}\,(\hat{\theta}|\theta) = [I(\theta)]^{-1}$$

En consecuencia, $I(\theta)$ puede utilizarse para establecer los intervalos confidenciales expuestos en el apartado anterior, donde:

$$S_e = \frac{1}{\sqrt{I(\theta)}}$$

si se prefiere:

$$S_e = [I(\theta)]^{-1/2}$$

S_e puede sustituirse por:

$$[I(\theta)]^{-1/2}$$

Ejemplo

Se aplicó un test de 25 ítems a una muestra de 1.000 personas. La función de información para $\theta = 2{,}5$ resultó ser $I(\theta) = 4$. Al nivel de confianza del 95%, ¿entre qué valores se estima que se encontrará θ?

1. NC 95%: $Z_c = \pm 1{,}96$.
2. Error máximo: $(Z_c)[I(\theta)]^{-1/2} = (1{,}96)[4]^{-1/2} = = (1{,}96)(1/\sqrt{4}) = 0{,}98$.
3. $(\hat{\theta} - 0{,}98) \leqslant \theta \leqslant (\hat{\theta} + 0{,}98)$
   $(2{,}5 - 0{,}98) \leqslant \theta \leqslant (2{,}5 + 0{,}98)$
   $1{,}52 \leqslant \theta \leqslant 3{,}48$.

La función de información (FI) es, pues, un indicador de la precisión del test. Tiene sentido conceptual la denominación de «función de información» dado que cuanto mayor sea $I(\theta)$, menor será el error típico de estimación y, por tanto, mayor será la información que las estimaciones aportan sobre el parámetro θ. Si la FI se calcula para todos los niveles de θ, se obtiene una curva del tipo de las que aparecen en la figura 7.21.

Según la figura 7.21, el test 1 aporta información máxima para los valores de θ en torno a 1,5, mientras que el test 2 tiene su eficacia máxima para

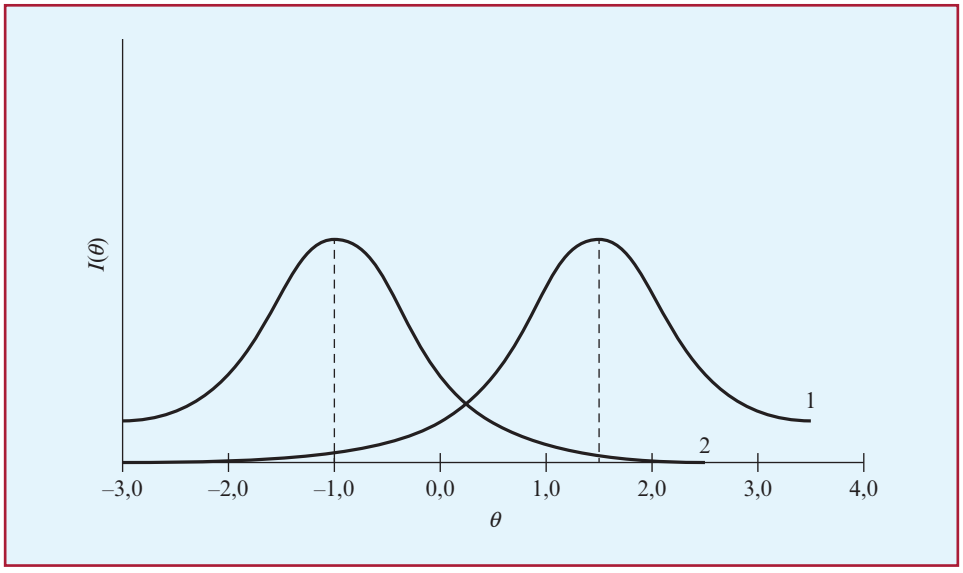


Figura 7.21.—Funciones de información de dos tests.

valores de θ en torno a –1. No se le escapará al lector la importancia de la FI a la hora de construir un test claramente se elegirá el 1 para medir en torno a $\theta = 1{,}5$ y el 2 para hacerlo en torno a $\theta = -1$. Además, si se dispone de un conjunto de ítems calibrados, esto es, con los parámetros estimados, se puede construir un test de tal manera que se ajuste a una FI previamente determinada según los objetivos perseguidos. Por ello, la FI se convierte en una herramienta imprescindible a la hora de construir y analizar test.

7.3. Funciones de información de los modelos logísticos

Vista la fórmula general de $I(\theta)$, vamos a especificar qué valores toma para cada uno de los modelos logísticos de uno, dos y tres parámetros, ya que la derivada de las CCI, $P_i'(\theta)$, no será la misma en los tres modelos. En el siguiente cuadro tomado de Hambleton y Swaminathan (1985) se ofrecen los valores de $P_i'(\theta)$ y de $I(\theta)$ para los tres modelos.

Modelos	$P_i'(\theta)$	$I(\theta)$
Un parámetro	$DP_i(\theta)Q_i(\theta)$	$\sum_{i=1}^{n} D^2 P_i(\theta)Q_i(\theta)$
Dos parámetros	$Da_iP_i(\theta)Q_i(\theta)$	$\sum_{i=1}^{n} D^2 a_i^2 P_i(\theta)Q_i(\theta)$
Tres parámetros	$\dfrac{Da_iQ_i(\theta)[P_i(\theta) - c_i]}{(1 - c_i)}$	$\sum_{i=1}^{n} \dfrac{D^2 a_i^2 Q_i(\theta)[P_i(\theta) - c_i]^2}{P_i(\theta)(1 - c_i)^2}$

7.4. Función de información de los ítems

Todos los conceptos anteriores referidos a la función de información del test son aplicables a cada ítem por separado. Precisamente una de las propiedades más importantes de la *función de información del test es que es la suma de las funciones de información de los ítems.* Análogamente a lo dicho para el test, la FI del ítem viene dada por:

$$I(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad [7.38]$$

donde $P_i'(\theta)$, $P_i(\theta)$ y $Q_i(\theta)$ tienen idéntico significado al ya señalado en el apartado anterior para el test. Nótese que la única diferencia con la Fl del test es que ha desaparecido el sumatorio, que indicaba que, para obtener la FI del test, había que sumar las FI de los ítems. Esta propiedad aditiva de la FI del test respecto de las de los ítems va a permitir poder confeccionar su forma según convenga eligiendo los ítems con una FI determinada.

La FI de los ítems constituye un poderoso instrumento para el análisis de los ítems, indicando no solo la cantidad de información que el ítem aporta a la medida de θ, sino también, y lo que es tal vez más importante, a qué nivel de θ aporta dicha información (véase lo dicho en la figura7.22).

El ítem 1 aporta información máxima en torno a valores de $\theta = -1{,}5$; el ítem 2, en torno a $\theta = 0$, y el ítem 3, para $\theta = 2$. Es importante advertir que si se está interesado en medir θ para valores bajos, por ejemplo, entre –2 y –1, el ítem 1 le daría mucha más información que el 2, y para valores altos el 3. Actualmente la FI de los ítems es el método de análisis de ítems más utilizado por los constructores de test, permitiéndoles mediante la combinación de los ítems obtener test ajustados a sus necesidades. Por ejemplo, si se lleva a cabo una selección de personal en la que se va a elegir a solo unos pocos muy competentes, se construiría un test formado por ítems del tipo del 3, que es el que más información aporta para niveles altos de θ. La FI también permitirá disminuir dramáticamente el número de ítems de un test sin pérdida relevante de la información aportada, descartándose aquellos que apenas aporten información a la medición.

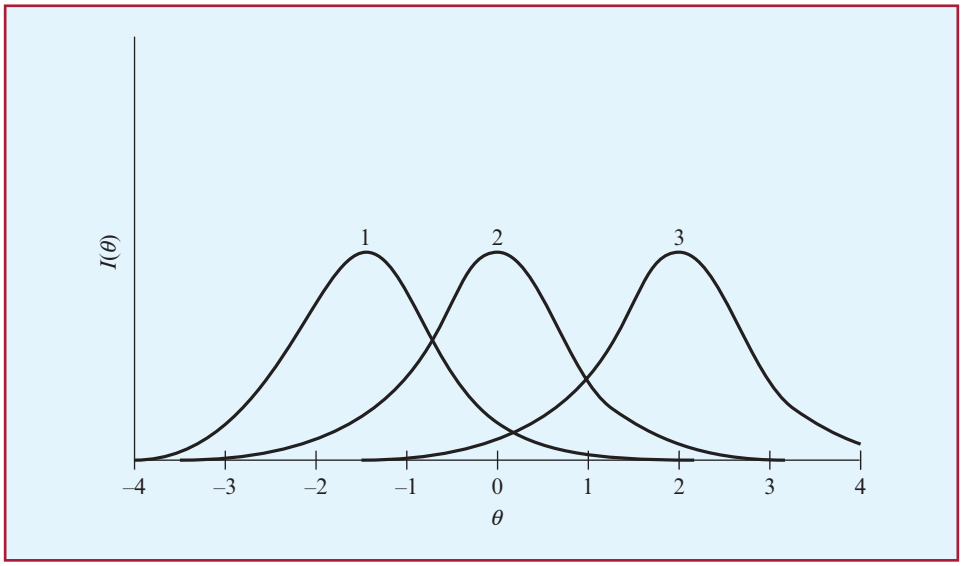


Figura 7.22.—Funciones de información de tres ítems.

La FI de los ítems para los modelos logísticos de uno, dos y tres parámetros es la misma que la ya explicitada para el test en el apartado anterior, prescindiendo del signo sumatorio.

En la figura 7.23 aparece representada la función de información de un test, obtenida al sumar las FI de sus cinco ítems.

7.5. Información máxima

En los *modelos logísticos de uno y dos parámetros* la información aportada por los ítems es *máxima para* $\theta = b$ y en el de *tres parámetros* para:

$$\theta = b + (1/Da)\{\ln[1/2 + (1/2)\sqrt{(1+8c)}]\} \qquad [7.39]$$

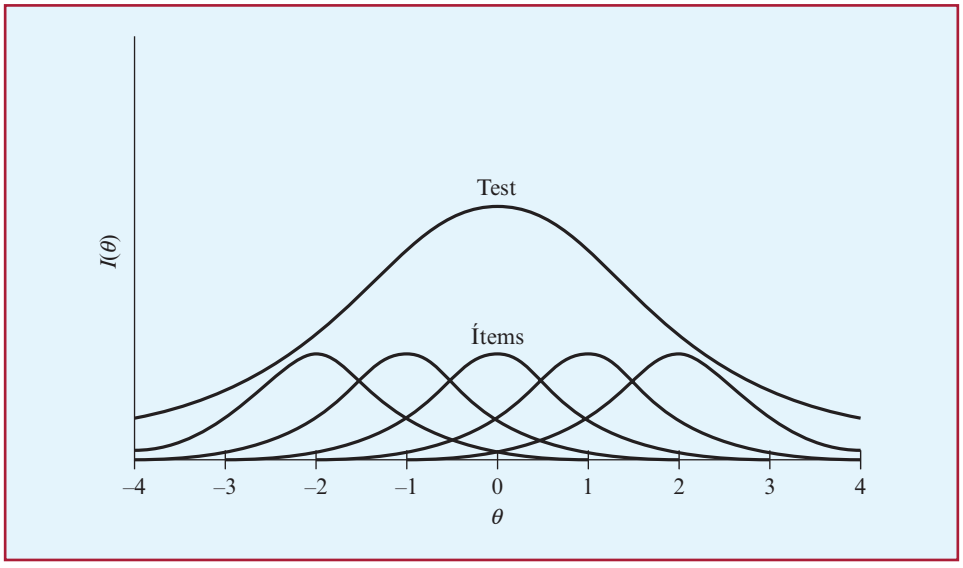


Figura 7.23.—Funciones de información de cinco ítems y función de información del test formado por esos cinco ítems.

La cuantía de la información aportada en ese punto de θ, en el que dicha información es máxima, viene dada por $D^2/4$ y $D^2a^2/4$ para los modelos logísticos de uno y dos parámetros respectivamente, y por

$$I(\theta) = [D^2a^2/8(1-c)^2]\,[1 - 20c - 8c^2 + (1+8c)^{3/2}] \qquad [7.40]$$

para el modelo de tres parámetros (Lord, 1980; Hambleton y Swaminathan, 1985).

EJEMPLO

Cierto ítem de un test se ajusta a un modelo logístico de tres parámetros con $a = 1$, $b = 0{,}75$ y $c = 0{,}20$. ¿Para qué valor de θ aporta más información la medición realizada por el ítem? ¿Cuál es la cuantía de la información aportada para dicho valor de θ? Sustituyendo en las fórmulas expuestas:

$$\theta = 0{,}75 + [(1/1{,}7)(1)]\{\ln[1/2 + 1/2\sqrt{1+8(0{,}20)}]\} = 0{,}91$$

$$I(\theta) = [(1{,}7)^2(1)^2/8(1-0{,}20)^2] \times \{1 - 20(0{,}20) - 8(0{,}20)^2 + [1+8(0{,}20)]^{3/2}\} = 0{,}49$$

Por tanto, el ítem aporta información máxima a la medida de θ cuando $\theta = 0{,}91$; sería un ítem óptimo para utilizarlo en la medición de θ para valores en torno a 0,91; en concreto para $\theta = 0{,}91$ la información aportada por el ítem es de 0,49.

7.6. Ponderación óptima de los ítems

Como señalan Birnbaum (1968) y Lord (1980), es posible asignar a los ítems ciertas ponderaciones para maximizar la información que proporcionan. Si a cada ítem se le asigna, por ejemplo, una ponderación w_i, la función de información del test vendrá dada por:

$$I(\theta, X) = \frac{\left[\sum_{i=1}^{n} w_i P_i'(\theta)\right]^2}{\sum_{i=1}^{n} w_i^2 P_i(\theta) Q_i(\theta)}$$

y el valor de las ponderaciones w_i que maximizan la función de información serán para los modelos logísticos:

Un parámetro: $w_i = D$
Dos parámetros: $w_i = Da_i$
Tres parámetros: $w_i = [Da_iP_i(\theta) - c_i]/[P_i(\theta)(1-c_i)]$

Ello quiere decir que en el *modelo logístico de un parámetro* los ítems *no* se ponderan distintos unos de otros, todos se multiplican por la constante D. La puntuación total del test (X) viene dada por la suma de las puntuaciones de los ítems, $\sum_{i=1}^{n} x_i$.

Multiplicada por D, es decir:

$$X = D\sum_{i=1}^{n} x_i \qquad [7.41]$$

En el *modelo logístico de dos parámetros* cada ítem se pondera multiplicándolo por su parámetro a_i; en consecuencia, la puntuación total del test (X) vendrá dada por:

$$X = D\sum_{i=1}^{n} a_i x_i \qquad [7.42]$$

donde n es el número de ítems y a_i el índice de discriminación de cada ítem.

Análogamente, para el *modelo logístico de tres parámetros* la puntuación total del test (X) vendrá dada por:

$$X = \sum_{i=1}^{n} \frac{Da_i[P_i(\theta) - c_i]}{P_i(\theta)(1-c_i)} x_i \qquad [7.43]$$

Al contrario que en los casos anteriores, las ponderaciones en el modelo de tres parámetros dependen del nivel de θ, reflejado en la fórmula por $P_i(\theta)$, característica esta no muy deseable para un modelo.

7.7. Eficiencia relativa de dos test

Una de las aplicaciones prácticas más interesantes de la FI es que permite comparar de un modo

muy adecuado la eficacia de dos test para medir θ a sus distintos niveles o valores. Se denomina *eficiencia relativa* (ER) de dos test para un determinado valor de θ al cociente entre las funciones de información de cada test para dicho valor de θ:

$$\text{ER} = \frac{I(\theta_x)}{I(\theta_y)} \qquad [7.44]$$

donde:

$I(\theta_x)$: Función de información del test X para $\theta = \theta_j$.
$I(\theta_y)$: Función de información del test Y para $\theta = \theta_j$.

Por ejemplo, si la FI de un test X para $\theta = 1$ vale $I(\theta_x) = 10$ y la FI de otro test Y para ese mismo valor de $\theta = 1$ vale $I(\theta_y) = 5$, la eficacia relativa en ese punto de θ vendrá dada por 10/5 = 2. A ese nivel ($\theta = 1$) el test X aporta el doble de información que el Y.

Ahora bien, tal vez ello no ocurra a todos los niveles de θ, denominándose *función de eficiencia* a la curva que une los valores de la eficiencia relativa calculados a los distintos niveles de θ.

Véase en la figura 7.24 la función de eficiencia para dos test, X e Y. Para valores de θ por debajo de –1,5 el test X es menos eficiente que el Y, ya que esos valores $I(\theta_x)/I(\theta_y) < 1$. Para valores de θ entre –1,5 y 2 ocurre lo contrario, el test X es más eficiente: $I(\theta_x)/I(\theta_y) > 1$. Finalmente, para valores de θ superiores a 2 la eficacia de ambos es similar. Trate el lector, a modo de ejercicio, de dibujar dos (posibles) funciones de información que podrían tener los dos test para dar lugar a la función de eficiencia aquí representada.

Las *aplicaciones* de la función de eficiencia son numerosas, amén de la obvia de comparar dos test, que no es poco; piénsese, por ejemplo, en la comparación de un test consigo mismo, pero cuyos ítems se ponderan de diferente modo en dos ocasiones, o se les cambia el número de alternativas, etc. La gran ventaja en todos los casos es que la función de eficiencia permite establecer estas comparaciones para los distintos niveles de θ. Por ejemplo, Lord (1980, p. 111) ilustra gráficamente cómo al disminuir el número de alternativas de los ítems de un test (SAT-V) este aumenta su eficiencia para las personas competentes (valores altos de θ), mientras disminuye para los niveles bajos de θ, lo que representa un nuevo enfoque al problema clásico del número óptimo de alternativas por ítem.

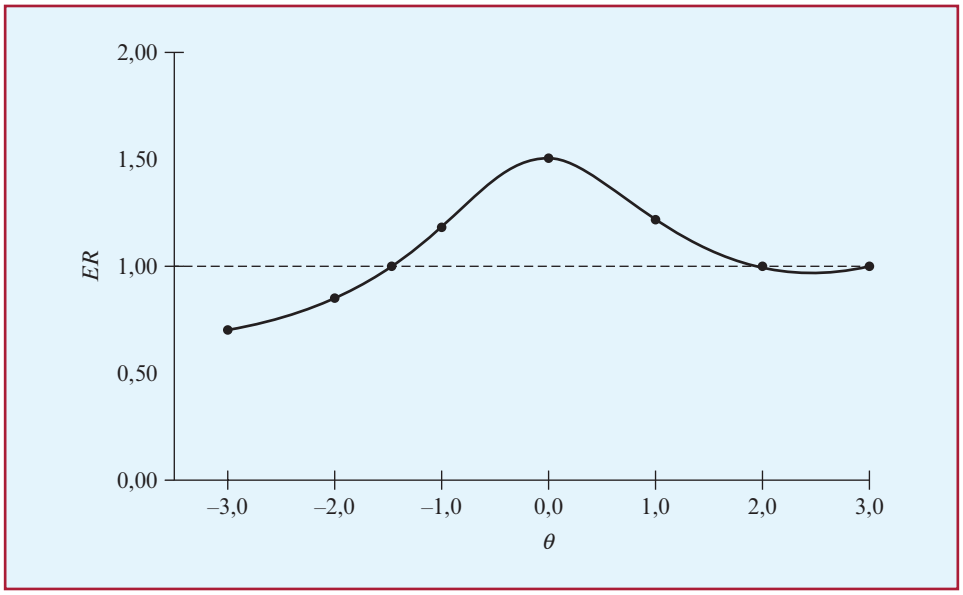


Figura 7.24.—Función de eficiencia de dos test.

7.8. Función de información y transformaciones de θ

La función de información es como se ha visto de gran utilidad en el estudio y análisis de los test y de los ítems, y es actualmente masivamente utilizada a tal efecto por los aplicadores, usuarios y constructores de test. Para su correcta interpretación ha de tenerse en cuenta que *la FI depende de la escala en la que se encuentren expresados los valores de* θ. Y, dado que esta escala es en cierto modo arbitraria, como ya hemos visto, el concepto y valor de la FI no es absoluto, depende de la escala elegida para θ. Esto no supone un grave inconveniente o desdoro para la utilidad de la FI, pero su ignorancia puede llevar a conclusiones erróneas acerca de la información proporcionada por el test para los diferentes niveles de θ. En concreto, según demuestra Lord (1980), si se lleva a cabo una transformación monotómica de θ, la FI queda dividida por el cuadrado de la derivada de la transformación. Si denominamos $I(\theta)$ a la FI original, θ_t a la transformación realizada e $I(\theta_t)$ a la FI resultante tras la transformación de θ en θ_t, lo dicho puede expresarse del siguiente modo:

$$I(\theta_t) = \frac{I(\theta)}{I(\theta'_t)^2} \qquad [7.45]$$

La nueva FI resultante $I(\theta_t)$ no solo puede ser completamente distinta a la original $I(\theta)$, sino que, como señala Lord (1980), *una FI puede transformarse en otra cualquiera con tal de que se elija la transformación de* θ *oportuna.*

En suma, la FI está íntimamente ligada a la métrica de θ y su interpretación y, en consecuencia, ha de ceñirse a ella. Sin embargo, *la eficiencia relativa de dos test no se ve afectada por la transformación monotómica de la escala de* θ. Veámoslo a continuación.

Para dos test X e Y su eficiencia relativa (ER) viene dada como se ha visto por:

$$\text{ER} = \frac{I(\theta_x)}{I(\theta_y)}$$

La ER tras la transformación (ER_t) será a su vez:

$$\text{ER}_t = \frac{I(\theta_{tx})}{I(\theta_{ty})} = \frac{I(\theta_x)/I(\theta'_{tx})^2}{I(\theta_y)/I(\theta'_{ty})^2}$$

Ahora bien, $(\theta'_{tx}\theta)^2 = (\theta'_{ty})^2$; puesto que ambos test están en la misma escala θ, la transformación realizada θ_t es idéntica en ambos, y así lo serán consecuentemente las derivadas correspondientes. Por tanto:

$$\text{ER}_t = \frac{I(\theta_{tx})}{I(\theta_{ty})} = \frac{I(\theta_x)}{I(\theta_y)} = \text{ER}$$

Esta propiedad de la invarianza de la ER respecto de transformaciones monotómicas de θ hace de la ER un excelente instrumento para el análisis comparativo de los test.

EJERCICIOS

1. En la tabla adjunta se ofrece la información aportada por los cinco ítems de un test para distintos valores de θ:

Ítem	θ						
	–3	**–2**	**–1**	**0**	**1**	**2**	**3**
1	0,01	0,03	0,06	0,07	0,24	0,08	0,02
2	0,02	0,06	0,10	0,20	0,30	0,10	0,07
3	0,07	0,15	0,20	0,15	0,09	0,04	0,01
4	0,20	0,60	0,30	0,20	0,10	0,03	0,02
5	0,01	0,06	0,09	0,10	0,20	0,70	0,22

1. Elabore la función de información del test para los valores θ de la tabla y represéntela gráficamente.
2. Si se construyen dos test, el A formado por los ítems 1, 3 y 4, y el B por los ítems 2 y 5, ¿cuál es la información dada por cada uno de ellos para $\theta = 2$? ¿Cuál es la eficiencia relativa de A y B para $\theta = 2$?
3. Si se desea disponer de un test que en el punto $\theta = 1$ ofrezca una información de 1,5 a base de ítems todos ellos idénticos al ítem 2, ¿cuántos ítems debería tener al menos el test?
4. Si los cinco ítems constituyesen un test referido al criterio destinado a dividir a las personas en expertos/no expertos, ¿a qué nivel de θ se llevaría a cabo el corte con una mayor precisión?
5. ¿Qué función de información debería tener un sexto ítem para que la función de información del nuevo test de seis ítems fuese la siguiente para los valores θ?: 0,40, 1,00, 1,02, 1,10, 1,20, 0,97, 0,40.

2. Mediante el modelo logístico de un parámetro se estimó que la dificultad de uno de los ítems de un test era $b = 1$.

1. Para el nivel $\theta = 2$,
   - 1.1. ¿Qué cantidad de información aporta este ítem a la medición?
   - 1.2. ¿A qué nivel de θ aporta máxima información el ítem?
   - 1.3. ¿Cuál es el valor máximo de información aportada por el ítem?

3. Utilizando el modelo logístico de dos parámetros, se estimaron los siguientes valores de los parámetros para uno de los ítems del test: $a = 0{,}50$ y $b = 1{,}00$.

1. Para un valor de $\theta = 2$,
   - 1.1. ¿Qué cantidad de información aporta este ítem a la medición?
   - 1.2. ¿A qué nivel de θ aporta máxima información el ítem?
   - 1.3. ¿Cuál es el valor máximo de información aportada por el ítem?

4. Mediante el modelo logístico de tres parámetros se estimaron los siguientes valores de los parámetros para uno de los ítems del test: $a = 0{,}50$, $b = 1{,}00$ y $\theta = 0{,}20$.

1. Para un valor de $\theta = 2$,
   - 1.1. ¿Qué cantidad de información aporta este ítem a la medición?
   - 1.2. ¿A qué nivel de θ aporta máxima información el ítem?
   - 1.3. ¿Cuál es el valor máximo de información aportada por el ítem?

5. A continuación aparecen los valores de los parámetros estimados para tres ítems de un test mediante el modelo logístico de tres parámetros.

Ítems	*a*	*b*	*c*
1	0,4	1,0	0,20
2	0,8	1,2	0,10
3	1,1	0,7	0,25

1. ¿Qué cantidad de información aporta cada uno de los ítems para el nivel de $\theta = 1{,}5$?
2. ¿Cuál es el valor de la función de información del test formado por esos tres ítems para $\theta = 1{,}5$?
3. Si hubiese que descartar uno de los ítems, tomando como criterio la información aportada, ¿cuál de los tres se eliminaría?

SOLUCIONES

1.1. 0,31, 0,90, 0,75, 0,72, 0,93, 0,95, 0,34.
2. A: 0,15; B: 0,80; ER = 0,19.
3. 5.
4. $\theta = +2$.
5. 0,09, 0,10, 0,27, 0,38, 0,27, 0,02, 0,06.

2.1.1. 0,39.

1.2. 1.
1.3. 0,72.
3.1.1. 0,15.
1.2. 1.
1.3. 0,18.
4.1.1. 0,11.
1.2. 1,32.
1.3. 0,12.
5. 1. 0,078; 0,374; 0,376.
2. 0,83.
3. 1.

8. BANCOS DE ÍTEMS

8.1. Concepto y desarrollo

Un banco de ítems no es otra cosa que un conjunto de ítems organizados cuyas propiedades psicométricas se conocen. La idea no es nueva, de una forma u otra siempre se han utilizado, pero en la actualidad se dan dos circunstancias potenciadoras. Por un lado, los ordenadores e internet ofrecen grandes ventajas a la hora de un almacenamiento eficaz y de una recuperación y búsqueda rápidas y eficientes. En este sentido, el ordenador no añade nada sustantivo desde el punto de vista conceptual, pero sus posibilidades instrumentales son notorias y obvias. Por otra parte, y más importante, la TRI permite, como se ha visto, expresar las propiedades de los ítems en términos de parámetros invariantes respecto de las personas, por lo que los profesionales e investigadores pueden elegir el tipo de ítem más indicado para sus objetivos, siempre, claro está, que el banco de ítems sea suficientemente amplio y heterogéneo. Prueba documental de esta vigencia es, por ejemplo, el número especial dedicado al tema por la revista *Applied Psychological Measurement* en 1986, y especialmente interesante el trabajo pionero de Chopin (1976), así como los de Millman y Arter (1984), Wright y Bell (1984), Ward y Murray (1994), Bergstrom y Gershon (1995) o Umar (1999). Buenas revisiones recientes pueden consultarse en Vale (2006) y Muckle (2016), y en español en Barbero (1996, 1999).

Dos aspectos de los bancos de ítems han de entenderse cabalmente: uno, cómo se construyen, y otro, cómo se procede a partir de ellos para elaborar test con determinadas características. Para construir un banco de ítems, una vez definido de forma adecuada el constructo que mide el banco, hay que desarrollar los ítems de forma pertinente (Haladyna y Rodríguez, 2013; Lane et al., 2016; Moreno, Martínez y Muñiz, 2006, 2015) y luego estimar los parámetros correspondientes, tal como se ha expuesto al presentar los modelos de TRI. Sería, valga la analogía, como construir un inmenso test bajo la óptica de la TRI, aunque nada impide, naturalmente, utilizar también los indicadores de la teoría clásica. Existe, no obstante, el problema adicional de cómo añadir nuevos ítems una vez que se han calibrado *n* ítems por el procedimiento anterior. Supóngase que se han calibrado inicialmente 500 ítems en una muestra de 20.000 personas; se dispone, por tanto, de un banco de 500 ítems. Si se elaboran 100 nuevos ítems, el problema de añadirlos al banco radica en que hay que calibrarlos en la misma métrica que el banco. Hay varios posibles diseños para llevar a cabo esta equiparación métrica. El más habitual y práctico consiste en aplicar los nuevos ítems a una muestra amplia de personas, a las que también se aplica otro test compuesto por ítems pertenecientes al banco ya calibrado (test de anclaje). Este test común de anclaje u otros diseños permite establecer la conexión entre las dos métricas, la del banco y la surgida en la nueva calibración. Para más detalles, véase el apartado siguiente sobre la equiparación de las puntuaciones. Por estos procedimientos se puede ir aumentando el número de ítems del banco y disponer de una buena descripción de la variable a medir. En realidad, todo lo dicho en la TRI respecto a las invarianzas de las medidas de los test tiene sentido cuando se dispone de un banco de ítems. Si el modelo se ajusta a los datos, ciertamente la medida es invariante respecto del subconjunto de ítems (test) del banco elegidos para obtenerla.

Las posibilidades para la elaboración de test a partir de los bancos de ítems son inmensas. Amén del usuario, que se encuentra con los ítems hechos y calibrados, con lo que ello supone en ahorros de todo tipo, lo más importante es que permiten confeccionar test con determinadas características espe-

cificadas a priori. Generalmente, esta especificación se hace mediante la función de información. El usuario escoge una determinada FI, que suele denominarse «función de información objetivo» y que dependerá del tipo de personas con las que se va a utilizar el test, eligiéndose *ad hoc* los ítems adecuados para generar dicha FI objetivo. Así, por ejemplo, si se va a evaluar a personas de alta competencia se elegirá como objetivo una FI que dé información máxima para valores elevados de θ, y los ítems incluidos en el test serán los que se ajusten a esas exigencias. Los programas informáticos suelen ofrecer la FI de cada ítem, por lo que se puede obtener la del test requerido combinando pertinentemente las de los ítems. La ventaja sobre la teoría clásica es notoria: según un tipo de personas u otro, se utilizará el test más adecuado y las mediciones obtenidas estarán, sin embargo, en la misma escala θ.

Véanse en la figura 7.25 las FI de tres test. La utilización de uno u otro dependerá de los objetivos del psicólogo: el test 1 discrimina eficientemente entre las personas inferiores en θ; el 2, entre las medias, y el 3, entre las superiores; las medidas proporcionadas por los tres están en la misma escala θ. Las implicaciones para la construcción de pruebas adecuadas serán enormes, por poner un ejemplo bien conocido en nuestro país, como es el caso del test de admisión para médicos internos residentes (MIR) y psicólogos internos residentes (PIR); piénsese lo descabellado que sería utilizar un test cuya FI fuese similar a la 1 o a la 2 de la figura 7.25, pues, dado que solo se admite una mínima proporción de los candidatos, es obligado que el test proporcione información máxima para los más competentes, que serán los admitidos.

En la actualidad se dispone de software abundante para el manejo de los bancos de ítems, que va desde Excel de Microsoft hasta sofisticados programas comerciales que funcionan *online*. Muckle (2016) ofrece una buena descripción de algunos de los programas:

TAO:	www.taotesting.com
Exam Studio:	www.zoomorphix.com
Questionmark:	www.questionmark.com
ExamDeveloper:	www.examdesign.com
Fast Test:	www.fasttestweb.com
ADE:	www.castleworldwide.com
AUTHORize:	www.certmanserv.com
pan:	www.panpowered.com
ProExam bank:	www.proexam.org
ITS:	www.testys.com

Los bancos de ítems constituyen cada día más el centro neurálgico de los procesos evaluativos, reclamando la interacción entre los actores que par-

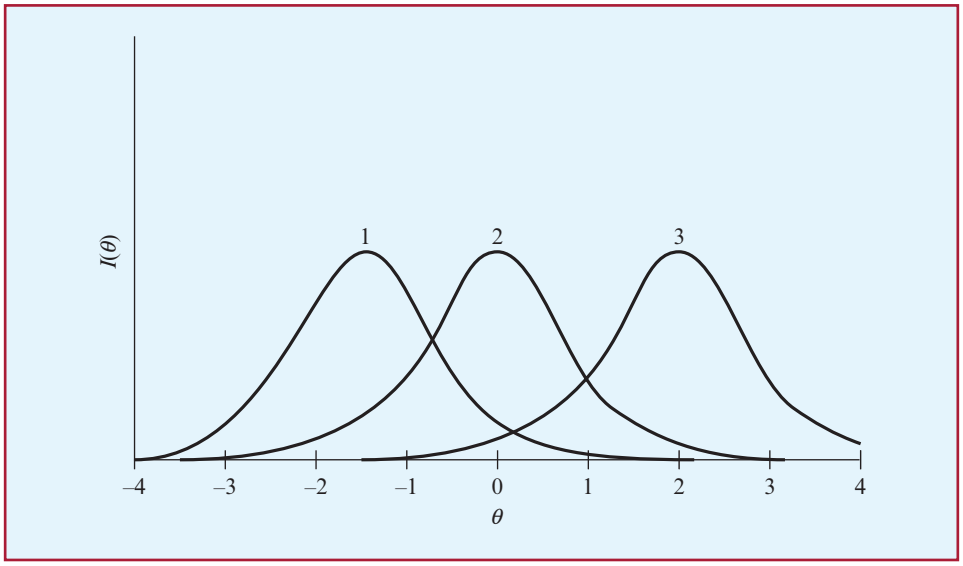


Figura 7.25.—Funciones de información de tres ítems.

ticipan en la evaluación: psicómetras, expertos en programación, especialistas en el constructo evaluado, autores de los ítems, los editores de test y otras partes legítimamente implicadas que habrá que determinar en cada situación concreta. El reto futuro de los bancos de ítems es dar acogida y gestionar de forma adecuada las grandes novedades que están ocurriendo en la evaluación psicológica y educativa, con especial mención para la construcción de ítems informatizados (Sireci y Zenisky, 2016), la generación automática de ítems (Gierl y Haladyna, 2013), los test adaptativos informatizados (Olea et al., 1999; Van der Linden y Glas, 2010) y el diseño óptimo de los test (Van der Linden, 2016). Una gran preocupación práctica de quienes gestionan los bancos es evitar perder ítems que se van «quemando» tras sucesivas aplicaciones, pues resulta muy costoso en tiempo y dinero incorporar nuevos ítems a los bancos, dado que es necesario todo un proceso de construcción, validación y calibración. Algunos expertos estiman que el precio de incorporar un nuevo ítem a un banco puede llegar a rondar los mil dólares (Downing, 2006; Muckle, 2016; Vale, 2006), lo cual da una idea de los medios necesarios, si tenemos en cuenta que un banco puede alcanzar varios miles de ítems.

9. EQUIPARACIÓN DE PUNTUACIONES

9.1. Concepto y técnicas

Seguramente el lector se hallará sorprendido de que, tras señalar repetidamente que la TRI permitía obtener medidas independientes de los instrumentos de medida, se esté hablando ahora de equiparar puntuaciones obtenidas con distintos instrumentos. La sorpresa está justificada si se ha comprendido todo lo anterior. Ciertamente, si se dispone de un conjunto de ítems (banco de ítems) calibrados, esto es, de los cuales se conocen sus parámetros previamente estimados, y el modelo de TRI se ajusta a los datos, entonces es indiferente qué subconjunto de ellos se utilice como test, pues todos darán las mismas estimaciones de θ para las personas. En otras palabras, el instrumento utilizado sí es invariante respecto de la medida de θ. Desde el punto de vista de la TRI no habría, por tanto, ninguna necesidad de establecer equivalencias entre los valores de θ dados por los distintos test, ya que están en la misma escala θ. (No confundir θ con las puntuaciones empíricas de las personas en los test o con las verdaderas estimadas en ellos.)

En conclusión, cuando se dispone de un banco de ítems calibrados la estimación de θ no depende del subconjunto de ítems (test) elegidos para estimarla; las estimaciones son invariantes respecto del instrumento de medida; por tanto, el establecimiento de equivalencias entre los test es innecesario dentro del marco de la TRI. Otra cosa bien distinta es que, por razones prácticas, se desee establecer una relación no entre los valores estimados de θ, que como se acaba de decir son los mismos, sino entre las puntuaciones verdaderas estimadas en cada test, o entre las empíricas. Dicha relación es inmediata a partir de la curva característica de los test, como se observa en la figura 7.26.

Supongamos que los test de la figura 7.26 constan de 100 ítems cada uno. Para un valor de $\theta = 0{,}30$ el valor de la CCT para el test *A* vale 50 y para el test *B* vale 10. Es decir, obtener 50 puntos en el test *A* es lo mismo que obtener 10 en el *B*, ambos generan una $\theta = 0{,}30$. Para el resto de los valores se procede análogamente, pudiendo establecerse una representación gráfica de la correspondencia entre las puntuaciones de uno y otro test. Estrictamente, 50 y 10 serían las puntuaciones verdaderas estimadas y, en la medida en que el modelo se ajuste a los datos, también las empíricas. Como el ajuste perfecto no es lo usual, existen algunos métodos para matizar la equiparación de las empíricas a partir de las verdaderas estimadas, en vez de considerarlas intercambiables sin más (véanse, por ejemplo, Hambleton y Swaminathan; 1985; Lord, 1980). No obstante, en la práctica las diferencias son insignificantes si el modelo se ajusta razonablemente (Lord y Wingersky, 1983), y, claro, si no se ajusta, hay que descartarlo. La única ventaja que puede tener hablar en términos de las puntuaciones en los test y no de los valores estimados de θ es la mejor comprensión por parte de los usuarios y clientes no muy familiarizados con la TRI, pero la información proporcionada es estrictamente la misma, ya que las puntuaciones estimadas verdaderas en el test son una transformación de θ mediante la curva característica del test.

Ahora bien, la situación real más habitual (otras muchas son pensables) no es la precedente e idílica en la que se dispone de un gran banco de ítems calibrados en la misma métrica del cual se

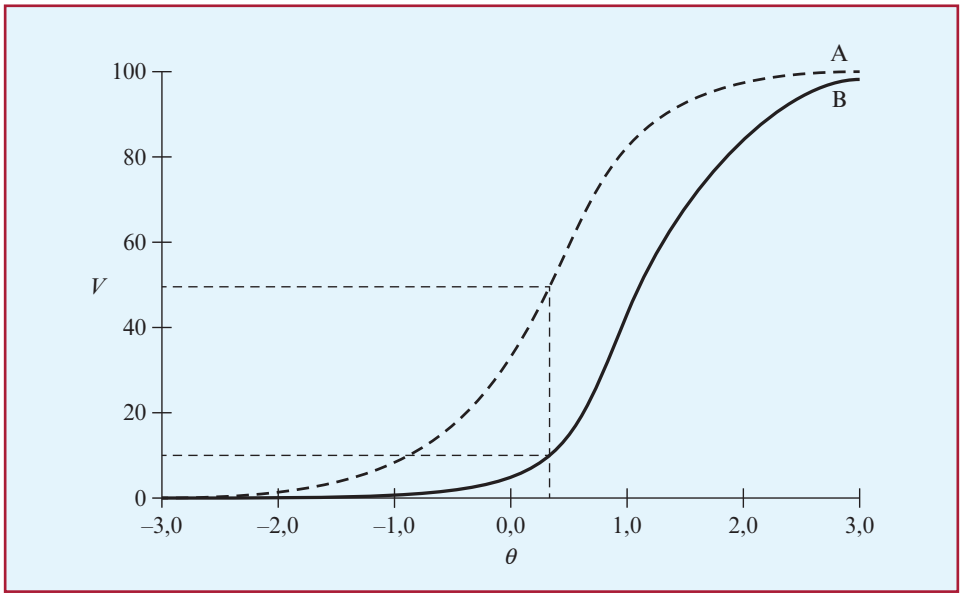


Figura 7.26.—Equiparación de las puntuaciones verdaderas estimadas de dos test, mediante sus curvas características.

extraen subconjuntos de ítems (test) para aplicar según las conveniencias. Entre otras cosas, no lo es porque los test se suelen «quemar» tras su utilización en una ocasión, especialmente en algunos países en los que existe una legislación que obliga a hacerlos públicos junto con las respuestas una vez que se han corregido y puntuado. Además, para calibrar los ítems e incluirlos en el banco habría que aplicarlos previamente a una muestra análoga a la de interés, con el consecuente riesgo de filtraciones. Piénsese que, en determinados casos, la puntuación puede ser de gran importancia para la vida futura de las personas: acceder o no a la universidad, ser admitido en una profesión, etc., con todo un negocio de escuelas y academias preparando y entrenando a tales efectos. Evidentemente, más que un problema teórico interesante de la medición psicológica o educativa, es este un problema típicamente aplicado y que se plantea sobre todo a las instituciones dedicadas a la construcción y uso sistemático de test con distintos fines. Así que la situación más frecuente en estos casos es disponer de un cierto banco de ítems calibrados en una métrica común y utilizados previamente y de un nuevo test sin calibrar en la misma métrica y que se va a usar en una próxima evaluación. El problema es ahora cómo equiparar las estimaciones de θ proporcionadas por el nuevo test aplicado a nuevas personas con las generadas por los anteriores test cuyos ítems están calibrados en otra métrica. En suma, es un problema de convertir la nueva métrica a la anterior. La dificultad proviene de que las personas, y los ítems, son distintos; si fuesen los mismos, sencillamente se estimarían los parámetros conjuntamente en una métrica común, con lo que los valores estimados de θ serían invariantes. La solución habitual es la ya citada en la construcción de los bancos de ítems, y consiste en incluir en el nuevo test un conjunto de ítems calibrados en la métrica del test (o banco) con el que se desea hacer equivaler el nuevo, esto es, utilizar un test de anclaje. Si, por ejemplo, el nuevo test consta de 100, se le añaden 20 calibrados que no van a ser generalmente utilizados en la calificación pero que servirán para establecer el vínculo entre la anterior y presente métrica. Para ello se calibra el nuevo test, incluidos los ítems de vínculo, estimando los parámetros de los ítems y los valores de θ para cada persona. A continuación se establece la relación lineal entre los parámetros b de los ítems comunes (test de anclaje) a ambas calibraciones, en el ejemplo 20, y que en términos generales vendrá dada por:

$$b_p = (K)b_n + D \qquad [7.46]$$

donde:

b_p: Valor del parámetro b de los ítems de anclaje en la métrica primitiva.
b_n: Valor en la nueva métrica.
K y D: Constantes a determinar según los datos.

La relación sería idéntica para los parámetros a; la única razón por la que suele utilizarse b para estimar K y D es su mayor estabilidad.

La relación entre los parámetros b, según se ha visto al tratar de las transformaciones de θ, es aplicable también a θ, con lo que las estimaciones del nuevo test se transforman a la métrica del antiguo. Los valores θ de las personas en la métrica primitiva (θ_p) estarán conectados con los de la nueva (θ_n) mediante la misma transformación lineal que b:

$$\theta_p = (K)\theta_n + D \qquad [7.47]$$

Para el caso del parámetro a:

$$a_p = \frac{a_n}{K}$$

El parámetro c no viene afectado por el cambio de métrica. Estimados b_p y b_n mediante la aplicación de un modelo de TRI, existen varios métodos para determinar el valor de las constantes K y D, entre los que cabe recomendar, según Hambleton y Swaminathan (1985), los de regresión, media-desviación típica, media-desviación típica robustas y métodos de curva característica. Aquí se ilustrará un ejemplo de equiparación, paso a paso, utilizando el método lineal de media-desviación típica para estimar las constantes K y D.

Aplicación

Veamos, mediante un ejemplo, las *etapas* que hay que seguir para equiparar las puntuaciones de dos test.

Sea un nuevo test X de coordinación visomotora que consta de 15 ítems, cinco de los cuales son comunes (test de anclaje) a otro test Y, también de coordinación visomotora, ya utilizado previamente.

Primero. Estimar los parámetros de los ítems y θ para el test primitivo Y mediante un modelo de TRI que se ajuste a los datos. (Si el test Y ya ha sido utilizado, es de suponer que se disponga de tales datos.)

Segundo. Estimar los parámetros de los ítems y θ para el nuevo test X, en el que además están incluidos los ítems de anclaje.

Tercero. Calcular las constantes K y D. (Supongamos que los valores estimados de b para los ítems de anclaje en la estimación primitiva hubiesen sido, respectivamente: 1,40, 0,70, 1,00, 1,20, 0,80 y en la nueva 1,24, 0,75, 0,90, 1,10, 0,74.)

Utilizando el método media-desviación típica las constantes K y D vienen dadas por:

$$K = \frac{S_{ya}}{S_{xa}} \qquad [7.48]$$

$$D = \bar{Y}_a - (K)\bar{X}_a \qquad [7.49]$$

donde S_{ya}, S_{xa}, $\bar{Y}_a$, $\bar{X}_a$ son las desviaciones típicas y medias de los ítems de anclaje en ambas calibraciones. En el presente ejemplo:

$$K = \frac{0{,}256}{0{,}196} = 1{,}306$$

$$D = 1{,}02 - (1{,}306)(0{,}946) = -0{,}215$$

Por tanto, la fórmula de equiparación viene dada por:

$$Y = (1{,}306)(X) - 0{,}215$$

Es decir, para equiparar las puntuaciones θ del nuevo test X a las del primitivo Y, han de multiplicarse por 1,306 y sumarles –0,215.

Nótese que las constantes K y D no son otra cosa que los valores provenientes de igualar las puntuaciones típicas de los valores de b en el test de anclaje para las dos calibraciones, la primitiva y la nueva:

$$\frac{X_a - \bar{X}_a}{S_{xa}} = \frac{Y_a - \bar{Y}_a}{S_{ya}}$$

despejando Y_a:

$$Y_a = \left(\frac{S_{ya}}{S_{xa}}\right)(X_a) - \left(\frac{S_{ya}}{S_{xa}}\right)(\bar{X}_a) + \bar{Y}_a$$

$$Y_a = \left(\frac{S_{ya}}{S_{xa}}\right)(X_a) + \left[\bar{Y}_a - \left(\frac{S_{ya}}{S_{xa}}\right)(\bar{X}_a)\right]$$

$$Y_a = (K)(X_a) + D$$

donde:

$$K = \frac{S_{ya}}{S_{xa}} \quad \text{y} \quad D = \bar{Y}_a - (K)(\bar{X}_a)$$

Cuarto. Convertir todos los valores del test X a la escala del Y, o viceversa, mediante la relación expuesta.

A continuación se ilustra lo dicho para los datos de un ejemplo. Sea un test (A) de inteligencia espacial compuesto de 20 ítems que se aplicó a una muestra de 600 estudiantes de bachillerato. Posteriormente, a otra muestra también de 600 estudiantes de bachillerato se aplicó otro test (B) de inteligencia espacial. En este segundo test B se incluyeron cinco ítems de anclaje utilizados también en el test A, los cuales aparecen en negrita en la tabla adjunta. Para la estimación de los parámetros se utilizó el modelo logístico de tres parámetros. En la tabla se reflejan los valores del parámetro de dificultad b estimado para todos los ítems.

Se van a transformar las puntuaciones del test B a la métrica del test primitivo A. Para calcular las constantes de equiparación K y D se obtiene la media y desviación típica de los ítems de anclaje para ambos test. Para el test A la media de los ítems de anclaje es 0,28 y la desviación típica 1,31; para el test B, la media es 0,12 y la desviación típica 1,06. Por tanto:

$$K = \frac{1{,}31}{1{,}06} = 1{,}24$$

$$D = 0{,}28 - (1{,}24)(0{,}12) = 0{,}13$$

De modo que, para transformar los valores de B en A se multiplica cada uno de ellos por 1,24 y se suma 0,13. Por ejemplo, el ítem 22, cuyo valor en el test B es 0,94, se transformaría en la nueva métrica en:

$$0{,}94(1{,}24) + 0{,}13 = 1{,}29$$

Ítems	Test A	Test B	Puntuaciones equiparadas
1	–0,79		–0,79
2	–0,59		–0,59
3	–1,07		–1,07
4	–1,35		–1,35
5	–1,57		–1,57
6	–0,33		–0,33
7	0,67		0,67
8	–0,07		0,07
9	0,61		0,61
10	–1,22		–1,22
11	–0,78		–0,78
12	0,37		0,37
13	–0,29		–0,29
14	0,06		0,06
15	–0,24		–0,24
16	**–1,90**	**–1,50**	–1,81
17	**–0,40**	**–0,50**	–0,44
18	**0,70**	**0,20**	0,54
19	**1,20**	**0,80**	1,16
20	**1,80**	**1,60**	1,95
21		–0,43	–0,40
22		0,94	1,29
23		0,12	0,28
24		1,64	2,16
25		0,97	1,33
26		0,98	1,34
27		0,61	0,89
28		–1,18	–1,33
29		–0,67	–0,70
30		–0,81	–0,87
31		–0,78	–0,84
32		–0,20	–0,11
33		–0,25	–0,18
34		–0,10	–0,01
35		–0,62	–0,64
36		–0,45	–0,43
37		0,63	0,91
38		0,38	0,60
39		–1,30	–1,48
40		0,09	0,24

Para el resto de los valores se procede del mismo modo, obteniéndose los valores de la columna de la derecha de la tabla, correspondiente a las puntuaciones del test *B* equiparadas para la métrica del test *A*. Nótese que, en el caso de los ítems de anclaje, se han promediado los valores asignados por la equiparación con los que de hecho tenían en el test *A*, dado que ambos valores por lo general no coincidirán exactamente. Así al ítem 16, por ejemplo, le correspondería un valor equiparado de: (–1,5)(1,24) + 0,13 = –1,73; sin embargo, en la columna de la tabla correspondiente a las puntuaciones equiparadas aparece un valor de –1,81 que proviene de promediar ambos valores: [–1,73 + + (–1,9)]/2 = –1,81. Aquí, en mor de la sencillez, solo se equiparan los valores de los índices de dificultad (*b*), pero la misma ecuación de equiparación se utilizaría para las puntuaciones de las personas. Los índices de discriminación (*a*) quedarían equiparados al dividirlos entre 1,24, y los valores del parámetro *c* serían comunes para ambas métricas. A modo de ejercicio puede el lector llevar a cabo la equiparación utilizando el modelo de regresión y comparar los resultados.

Señalar, finalmente, que, aunque aquí se ha ilustrado la equiparación mediante el método de media-desviación típica, debido a su sencillez, diversos autores (Hambleton y Swaminathan, 1985; Stocking y Lord, 1983) recomiendan el de la media-desviación típica robustas o el método de la curva característica (Haebara, 1980; Stocking y Lord, 1983); la lógica de este último consiste en estimar *K* y *D* de tal guisa que minimicen las diferencias entre las puntuaciones verdaderas estimadas por los test entre los que se trata de establecer la equivalencia.

En la práctica, afortunadamente, los programas informáticos utilizados con los modelos de TRI permiten llevar a cabo de un modo menos aparatoso el establecimiento de equivalencias entre dos test para la situación anteriormente descrita. El diseño ilustrado, en el que se utiliza un test común de anclaje entre los dos test a equiparar, es el más habitual en la práctica, por adaptarse bien a las situaciones «realmente existentes»; ahora bien, el diseño de equiparación más natural sería aplicar los dos test a equiparar a la misma muestra de personas y estimar los parámetros de ambos conjuntamente en la misma métrica. Cuando ello sea posible, así se recomienda. Incluso un tercer diseño (dos grupos equivalentes) consistiría en aplicar ambos test a dos muestras aleatorias de la misma población, en cuyo caso, dada la aleatoriedad, cabría esperar distribuciones parejas y se equipararían biunívocamente los valores de θ obtenidos en ambas calibraciones. Este método es altamente peligroso y no recomendable, amén de que para tal viaje no hace falta recurrir a la TRI, se puede llevar a cabo desde la óptica clásica, como de hecho se ha venido haciendo con cierta frecuencia. Para un tratamiento exhaustivo de los problemas y métodos implicados en la equiparación, véanse Von Davier (2011) o Kolen y Brennan (2014), y en español Navas (1996).

EJERCICIOS

1. Se aplicó un test de rendimiento académico de 80 ítems a una muestra aleatoria de 1.000 estudiantes de bachillerato. Posteriormente, a otra muestra aleatoria, también de estudiantes de bachillerato, se les aplicó otro test paralelo de rendimiento que incluía cinco de los ítems utilizados con la primera muestra. En ambos casos se estimaron los parámetros de los ítems y las puntuaciones θ de los estudiantes mediante el modelo logístico de tres parámetros.

En la tabla adjunta aparecen los valores de los parámetros de dificultad (*b*) estimados en las dos ocasiones.

Primera aplicación (P)	Segunda aplicación (N)
–1,5	–1,9
–0,5	–0,8
0,4	0,0
0,8	0,3
1,6	0,2

1. Calcule las ecuaciones que transforman los valores de los parámetros *a*, *b*, *c* y θ de la nueva métrica (segunda aplicación) en los de la métrica primitiva (primera aplicación).

2. Un estudiante obtuvo una puntuación $\theta = 1{,}3$ en la segunda aplicación, ¿qué puntuación θ le corresponde en el test utilizado con la primera muestra?
3. Los parámetros estimados en la primera aplicación para uno de los ítems fueron: $a = 0{,}75$, $b = 1{,}1$ y $c = 0{,}22$. ¿Qué parámetros le corresponderían en la nueva métrica?
4. Calcule la correlación entre las estimaciones de los parámetros b de los ítems de anclaje en las dos ocasiones.
5. ¿Cuál es la correlación entre los valores de b en la nueva métrica y las estimaciones hechas a partir de ellos para la métrica primitiva? Comente las diferencias entre este valor de la correlación y el hallado en el apartado anterior.

SOLUCIONES

1.1. $a_p = a_n/1{,}02$.
$b_p = 1{,}02(b_n) + 0{,}4$.
$c_p = c_n$.
$\theta_p = 1{,}02(\theta_n) + 0{,}4$.

2. 1,73.
3. 0,76, 0,69, 0,22.
4. 0,99.
5. 1.

10. FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS

10.1. Concepto

El concepto de funcionamiento diferencial de los ítems (FDI) ya se expuso al tratar el enfoque clásico. Como se indicaba entonces, se trata de asegurarse de que los ítems de un test funcionan de igual modo para dos o más grupos de personas evaluadas, no perjudicando ni favoreciendo a las personas pertenecientes a distintos grupos, por ejemplo hombres y mujeres, distintas culturas, nacionalidades, etc. En términos de la TRI, un ítem mostrará FDI para dos o más grupos si a valores iguales de θ no corresponden valores iguales de $P(\theta)$ en las curvas características de los grupos considerados. En la figura 7.27 aparece la curva característica de

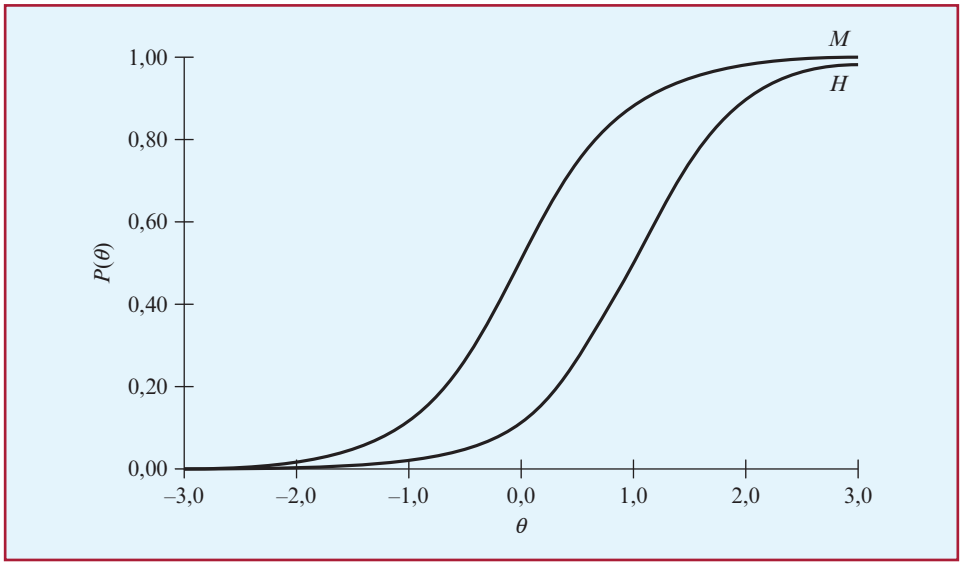


Figura 7.27.—Curva característica de un ítem en una muestra de mujeres (M) y en otra de hombres (H).

un determinado ítem para dos grupos, uno de hombres y otro de mujeres.

Nótese que para un mismo valor de $\theta = \theta_j$, la probabilidad de acertar este ítem es superior para las mujeres que para los hombres. Por tanto, el ítem funciona diferente para hombres y mujeres, estando claramente sesgado *contra* los hombres, es decir, a niveles iguales de competencia en la variable medida θ no se corresponden probabilidades iguales de superar el ítem, están sistemáticamente a favor de las mujeres. Más adelante se expondrán algunas técnicas para la evaluación y cuantificación del funcionamiento diferencial de los ítems en el marco de la TRI.

Como es fácil de entender, el problema del FDI viene acompañado de serias implicaciones en el uso de los test, pues, de darse tal sesgo, ciertos grupos sociales, clásicamente blancos-negros, mujeres-hombres, pobres-ricos, rurales-urbanos, etc. —cualquier otra partición es posible—, sufrirán las consecuencias. Si se toma una postura socialmente militante y se afirma de antemano que las variables psicológicas medidas han de tomar los mismos valores para los grupos citados, u otros, entonces la definición de sesgo adoptada es mucho más lasa, a saber, se hablará de sesgo siempre que se encuentren diferencias entre los grupos. Nótese que de la definición original no se deriva esto: las comparaciones no se establecen entre los grupos considerados globalmente, sino entre las personas de ambos grupos que tienen el mismo nivel en la variable medida. Es importante entender esta diferencia, pues es perfectamente posible que un ítem esté sesgado «contra» determinada subpoblación según el primer concepto de sesgo y, sin embargo, este mismo grupo obtenga puntuaciones superiores a la subpoblación «favorecida» por el sesgo. Dicha situación queda ilustrada en la figura 7.28, en la que aparece la curva característica de un ítem para dos grupos, hombres (H) y mujeres (M).

El ítem en cuestión está sesgado contra los hombres, pues para un mismo valor de θ los valores de $P(\theta)$ son inferiores si la persona es hombre. Sin embargo, la puntuación media de los hombres es superior a la de las mujeres, lo cual depende de las distribuciones de θ y no de las CCI. En la práctica, no se encontrarán habitualmente situaciones tan claras como la utilizada aquí con el fin de ilustrar las diferencias entre ambos enfoques.

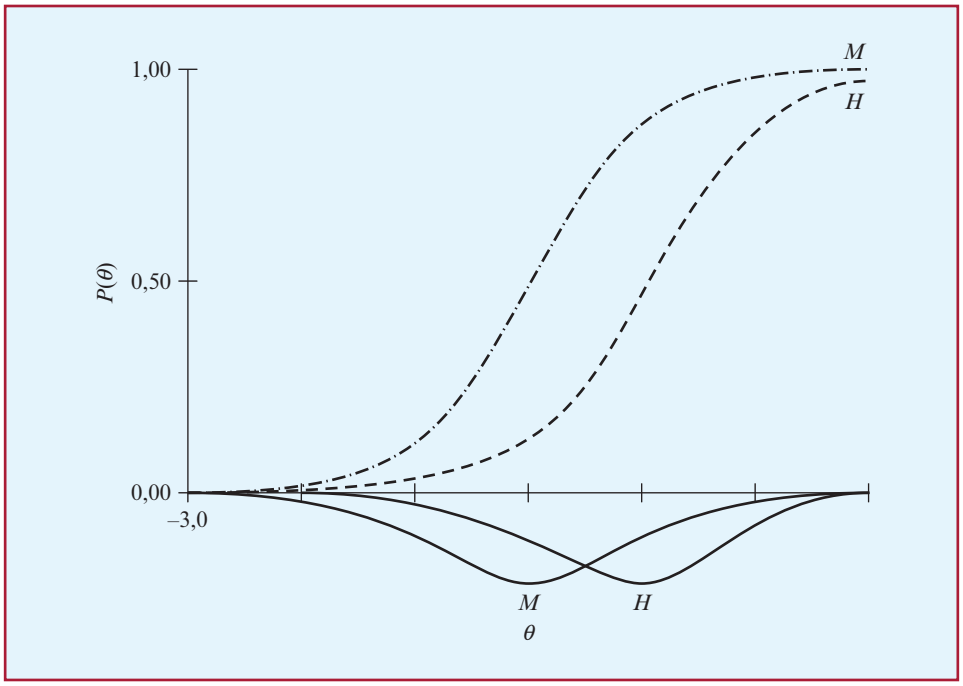


Figura 7.28.—Ítem con un funcionamiento diferencial para hombres y mujeres.

La psicometría se ocupa del sesgo tal como se definió en primer lugar, es decir, entiende que un ítem o un test están sesgados si personas igualmente competentes y pertenecientes a distintas subpoblaciones no tienen la misma probabilidad de superar el ítem (o test). Ahora bien, si dos personas tienen el mismo nivel en una variable, ¿a qué se puede deber que un ítem diseñado para medir esa variable pueda estar sesgado, esto es, pueda ser más favorable a uno que a otro? Las fuentes del sesgo son numerosas, y vienen generadas principalmente por el distinto bagaje cultural, social, económico, etc., de las personas. Dado que estos antecedentes históricos de las personas nunca serán los mismos, y pueden ser marcadamente distintos según la subcultura, si un ítem, o instrumento, en general, se apoya más en la de unos que en la de otros, tendrá altísimas probabilidades de no ser equitativo, de estar sesgado. Si, por ejemplo, un test de matemáticas está formulado de tal modo que exige un alto nivel de comprensión verbal, estará sesgado contra los lectores menos eficientes. En términos de diseño se confunde el efecto de la comprensión verbal con el de la competencia matemática, es decir, si una persona puntúa bajo en el test no sabremos a ciencia cierta si atribuirlo a su bajo rendimiento en matemáticas o a que su competencia verbal es limitada y no ha llegado a captar los problemas planteados. La casuística es interminable y puede decirse que estrictamente no existen pruebas exentas completamente de sesgo; más bien se trata de detectar la cantidad de sesgo tolerable. Expuesto brevemente el concepto de funcionamiento diferencial, véanse Shepard (1982), Holland y Wainer (1993), Osterlind y Everson (2009) o Dorans y Cook (2016) para un análisis detallado. En español Fidalgo (1996) y Gómez, Hidalgo y Gilera (2010) llevan a cabo muy buenas exposiciones del tema. A modo de ilustración, se describen a continuación algunas de las técnicas sencillas de que se valen los psicómetras para la detección del funcionamiento diferencial de los ítems. En la actualidad los programas informáticos utilizados para estimar los parámetros de la TRI permiten asimismo la evaluación del FDI.

10.2. Evaluación

Seguramente el método más eficiente para evitar en lo posible el funcionamiento diferencial de los ítems sea un cuidadoso análisis de su contenido por parte de varios expertos previo a su publicación. Una buena exposición sobre el modo de sistematizar y formalizar esta revisión es la de Tittle (1982). Hecha tal revisión y aplicados los ítems a las personas, aún cabe llevar a cabo ciertos análisis estadísticos que permiten detectar el funcionamiento diferencial en ítems escapados al análisis previo. A este tipo de técnicas estadísticas a posteriori nos referimos aquí, pero dejando claro que solo son un complemento de un escrutinio riguroso previo.

La TRI parece venir como anillo al dedo para la evaluación del FDI, como ya se apuntó al introducir el concepto. *Bajo la óptica de la TRI un ítem estará sesgado si su CCI no es la misma para los grupos en consideración.* En consecuencia, la lógica general para evaluar el FDI será estimar las CCI para ambos grupos y compararlas.

En las dos figuras que siguen se ilustran dos tipos de FDI. En la figura 7.29 el ítem representado está claramente sesgado «contra» los hombres, mientras que en la figura 7.30 el sesgo depende del nivel de θ: para valores bajos de θ el ítem está sesgado contra los hombres, y para niveles altos, contra las mujeres. En el primer caso hablamos de FDI uniforme, y en el segundo, de no uniforme, pues depende de los valores de θ.

Métodos para llevar a cabo la comparación de las CCI se han propuesto varios, que van desde una simple inspección visual de las CCI hasta complejos análisis estadísticos (Berk, 1982; Camilli y Shepard, 1994; Fidalgo, 1996; Hambleton y Swaminathan, 1985; Lord, 1980; Rosenbaum, 1987; Rudner et al., 1980; Shepard et al., 1981, 1984, 1985).

Aquí se comentarán brevemente algunos de ellos.

Método de las áreas

Por este método (Rudner, 1977; Rudner et al., 1980), en un primer paso, se estiman las CCI de los ítems cuyo funcionamiento diferencial se está estudiando para las dos (o más) subpoblaciones de interés, y a continuación se calcula el área comprendida entre las CCI. La cuantía del área constituye un índice de la discrepancia entre las CCI y, en consecuencia, del FDI, pues si ambas CCI coincidiesen, el área entre ambas sería cero, sería la misma CCI, no habría funcionamiento diferencial.

Más específicamente los *pasos* a seguir podrían concretarse en los siguientes:

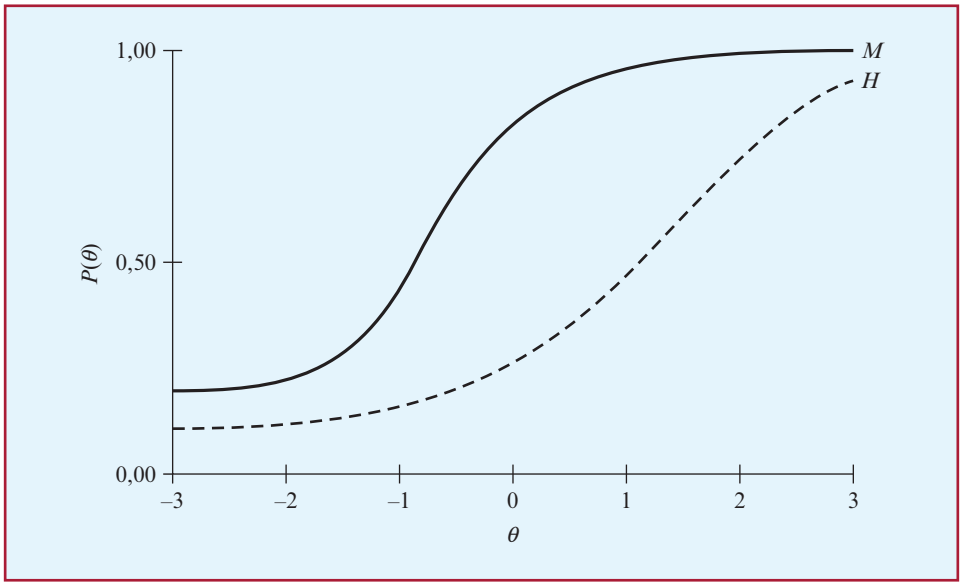


Figura 7.29.—Ítem con funcionamiento diferencial uniforme.

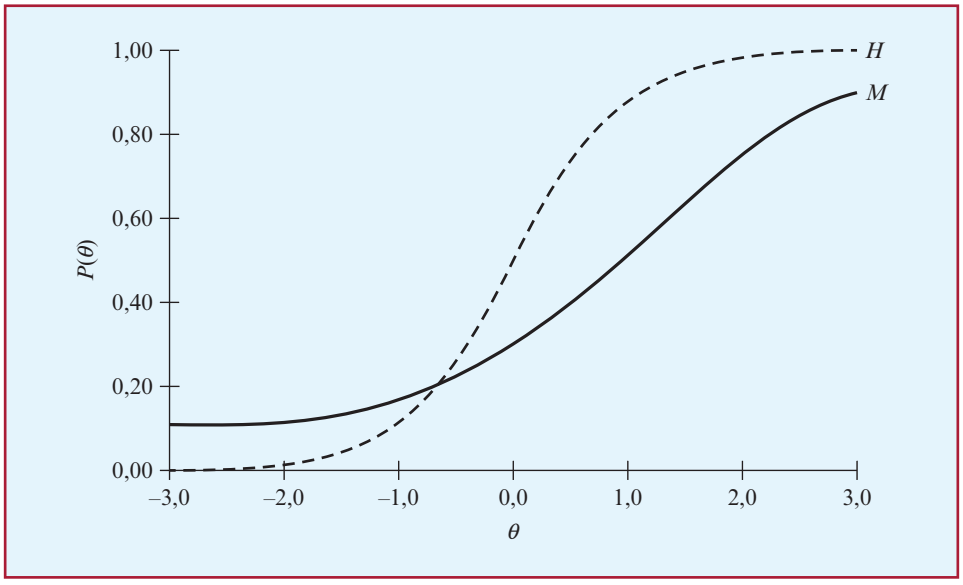


Figura 7.30.—Ítem con funcionamiento diferencial no uniforme.

1. Definir las subpoblaciones de interés, por ejemplo, mujeres-hombres, rural-urbano, universitarios-no universitarios, etc.

   El grupo más amplio suele denominarse «referencia», y el minoritario, «focal».
2. Elegir el modelo de TRI a utilizar.
3. Estimar los parámetros de los ítems para el modelo elegido en cada subpoblación. Los parámetros han de estar en la misma métrica en ambas subpoblaciones.
4. Se calcula el área comprendida entre ambas CCI:

$$A = \sum_{\theta=-4}^{\theta=+4} \left| P_R(\theta_j) - P_F(\theta_j) \right| \Delta\theta \qquad [7.50]$$

Se eligen los valores de θ entre –4 y +4 porque de hecho recogen la práctica totalidad de los valores empíricos, aunque en puridad $-\infty$ y $+\infty$ sería lo correcto. El cálculo del área será más exacto para valores pequeños del incremento de $\theta(\Delta\theta)$, siendo lo más usual utilizar $\Delta\theta = 0{,}005$. $P_R(\theta)$ y $P_F(\theta_j)$ son, respectivamente, los valores de $P(\theta)$ para la CCI en ambas subpoblaciones al ir sustituyendo θ desde –4 hasta +4 con incrementos sucesivos de 0,005. Es decir, para valores de θ = –4, –3,995, –3,990, –3,895, –3,890, ..., +3,995, +4. Nótese que la expresión situada a la derecha del sumatorio representa el área de un rectángulo de base $\Delta\theta = 0{,}005$ y altura $|P_R(\theta_j) - P_F(\theta)_j|$.

El método de Rudner no nos dice cuán grande ha de ser el área entre las CCI para decidir que hay FDI, es un método descriptivo, la decisión es subjetiva. Una idea comparativa puede obtenerse considerando varios ítems conjuntamente. Una ligera modificación al método de Rudner (1977) es la propuesta por Linn et al. (1981):

$$A = \sum_{\theta=-3}^{\theta=+3} \sqrt{[P_R(\theta_j) - P_F(\theta_j)]^2 \Delta\theta} \qquad [7.51]$$

Los autores han sugerido muy atinadamente que la importancia del FDI no será la misma para todos los niveles de θ, dependiendo de los objetivos perseguidos. Por ejemplo, si un test está destinado a la selección de un escaso número de aspirantes entre los más competentes, el FDI para niveles medios-bajos de θ será poco relevante. Los propios autores proponen algunos métodos de ponderación.

Raju (1988) ha propuesto una fórmula para calcular el área entre las CCI para los modelos logísticos de uno, dos y tres parámetros:

$$A = (1 - c) \left| \frac{2(a_2 - a_1)}{D a_1 a_2} \ln\left[1 + e^{D a_1 a_2 (b_2 - b_1)/(a_2 - a_1)}\right] - (b_2 - b_1) \right| \qquad [7.52]$$

donde a, b y c son los parámetros de los ítems; D, una constante que toma el valor 1,7, y θ = 2,7182, la base de los logaritmos neperianos. Para poder utilizar esta fórmula se asume que el valor del parámetro c es el mismo para los dos grupos analizados. Nótese que para el caso del modelo de dos parámetros desaparece el término c, y para el caso del modelo de un parámetro la fórmula se reduce a la diferencia absoluta entre los valores de los parámetros b de ambos grupos. Raju derivó posteriormente (Raju, 1990) una fórmula para el error típico del área, asumiendo que los cocientes entre el área y el error típico se distribuyen según la curva normal. Bajo este supuesto puede calcularse la significación estadística del área, siempre en el supuesto de que los parámetros c de ambas CCI son iguales. Las fórmulas anteriores para el cálculo del área entre las curvas características pueden expresarse conceptualmente de forma más general mediante integrales:

$$A = \int [P_R(\theta) - P_F(\theta)] d\theta \qquad [7.53]$$

Como se observa en la fórmula, las probabilidades del grupo de referencia se restan del focal; por tanto, si el grupo de referencia es consistentemente superior al focal a lo largo de la escala, el valor de A será positivo, y si ocurre lo contrario, será negativo. Si el FDI es no uniforme podría ocurrir que los valores positivos y negativos se anulasen, dando este índice la falsa impresión de que no existe FDI cuando en realidad sí lo hay, aunque no uniforme. Para evitar este inconveniente puede utilizarse la misma fórmula, pero elevando las diferencias al cuadrado y extrayendo la raíz cuadrada del resultado:

$$A = \sqrt{\int [P_R(\theta) - P_F(\theta)]^2 d\theta} \qquad [7.54]$$

La gran ventaja de los métodos de las áreas es su sencillez y que resultan muy intuitivos como forma de cuantificar el área observable entre las CCI del grupo de referencia y el focal. Tienen, sin embargo, dos claros inconvenientes. El primero, ya citado, es que dan el mismo peso a todas las superfi-

cies a lo largo del continuo de la variable medida θ, si bien se han propuesto algunas formas de ponderación, tales como el número de personas del grupo focal en cada rango de θ o incluso el número total de personas. El otro inconveniente es que si los valores del parámetro θ del ítem para el grupo focal y el de referencia son distintos, el valor del área resulta infinito, razón por la cual se acota el valor de los índices entre dos valores tales como (–3, +3) o (–4, +4), que en la práctica puede decirse que abarcan todos los valores empíricos, aunque teóricamente el rango vaya de $-\infty$ a $+\infty$.

Una vez calculada el área entre las CCI, cabe preguntarse a partir de qué valor del área se debe considerar que el ítem está funcionando diferencialmente para el grupo de referencia y el focal. No existen pruebas de significación estadística apropiadas, por lo que hay que proceder de forma descriptiva. Por ejemplo, puede tomarse como línea base la media de las áreas de todos los ítems y considerar con FDI aquellos que más se alejen de ese valor medio. O bien pasar a revisar directamente aquellos ítems cuyas áreas sean mayores. Una forma más rigurosa de proceder, aunque más costosa, es comparar los valores de las áreas obtenidas para los grupos de referencia y focal con aquellas estimadas a partir de submuestras aleatorias dentro del grupo de referencia y del focal, que representarían las áreas intercurva halladas por mero azar (Hambleton y Rogers, 1989). Lo más habitual es dividir aleatoriamente el grupo en dos mitades. Un posible criterio para decidir que un ítem tiene FDI sería que su área superase el máximo valor hallado al azar en las submuestras utilizadas. Una alternativa al uso de submuestras es la utilización de datos simulados, pues el empleo de submuestras tiene el inconveniente de que rebaja a la mitad el número de personas a partir de las cuales se estiman los parámetros, con la consiguiente inexactitud que ello introduce (Rogers y Hambleton, 1989). Aunque la determinación estadística precisa de esta cuantía constituya un problema metodológico interesante para los especialistas, en la práctica no supone un gran inconveniente, pues, ante la duda de si un ítem presenta FDI o no, lo aconsejable es revisarlo, dado que en este contexto más vale incrementar el número de errores de tipo I (revisar ítems que no presentan FDI) que de tipo II (no revisar ítems que presentan FDI).

Método de las probabilidades

Una forma de evitar los inconvenientes de ponderación de los métodos de las áreas es utilizar las diferencias de probabilidades entre el grupo de referencia y el grupo focal solo en las zonas de θ donde halla empíricamente personas del grupo focal. Basándose en las recomendaciones al respecto de autores como Linn y Harnisch (1981) y Shepard et al. (1984), Camilli y Shepard (1994) proponen dos índices sencillos y muy aconsejables. Nótese que, al utilizar únicamente los valores de θ para los cuales hay empíricamente personas del grupo focal, lo que se pretende es dar más importancia al FDI allí donde realmente hay personas pertenecientes al grupo focal, y no en otras zonas de θ donde no las hay. Veamos los dos índices de diferencias entre probabilidades (*DP*):

$$DP = \sum_{j=1}^{n_F} \frac{[P_R(\theta_j) - P_F(\theta_j)]}{n_F} \qquad [7.55]$$

donde:

- $P_R(\theta_j)$: Es la probabilidad que las personas del grupo de referencia tienen de superar el ítem para el valor θ_j. Este valor viene dado por la CCI del grupo de referencia.
- $P_F(\theta_j)$: Es la probabilidad que las personas del grupo focal tienen de superar el ítem para el valor θ_j. Este valor viene dado por la CCI del grupo focal.
- n_F: Es el número de personas del grupo focal.

El sumatorio va desde 1 hasta el número total de personas del grupo focal (n_F), es decir, solo se consideran aquellos valores de θ obtenidos por los miembros del grupo focal, lo cual, como ya se ha señalado, es una forma de autoponderación basada en las personas del grupo focal, dando más peso a las zonas de θ donde estas se encuentran y omitiendo las zonas de θ en las que no hay personas del grupo focal. No obstante, el método en sí no impide que la ponderación se haga utilizando el número de personas de la muestra total, opción que en al-

gunos casos podría interesar. Los pasos a seguir para calcular el índice *DP* pueden sintetizarse del siguiente modo:

1. Estimar por separado las CCI para el grupo de referencia y el focal.
2. Estimar las puntuaciones θ de las personas del grupo focal.
3. Calcular los valores de $P(\theta)$ para el grupo de referencia $[P_R(\theta)]$ y focal $[P_F(\theta)]$, correspondientes a los valores de las personas del grupo focal.
4. Hallar las diferencias $[P_R(\theta_j) - P_F(\theta_j)]$.
5. Se suman todas las diferencias $[P_R(\theta_j) - P_F(\theta_j)]$ y se divide el resultado entre el número de personas del grupo focal (n_F).

En el caso ideal de que no existiese FDI, los valores de $P_R(\theta_j)$ y de $P_F(\theta_j)$ coincidirían para todos los valores de θ_j, por lo que el valor del índice *DP* sería cero. El FDI aumentará a medida que *DP* se aleja de cero, bien positiva o negativamente. Si el valor es positivo, quiere decir que los valores $P_R(\theta_j)$ son superiores a los valores $P_F(\theta_j)$, lo que indicaría que el ítem en cuestión está perjudicando al grupo focal. Por el contrario, si el valor de *DP* es negativo, indica que el ítem perjudica al grupo de referencia. En el caso de FDI no uniforme el valor de *DP* podría ser engañoso, como ocurría con los métodos de las áreas, pues los valores negativos y los positivos podrían anularse, dando la falsa impresión de ausencia de FDI, cuando en realidad sí lo hay. Lo más sencillo y recomendable es contrastar los valores numéricos con los gráficos correspondientes, en los cuales puede observarse la naturaleza del FDI.

Ejemplo

Supongamos una muestra de 15 niños, 10 de los cuales son de origen urbano (grupo de referencia) y cinco de origen rural (grupo focal). En la tabla adjunta aparecen sus probabilidades de superar uno de los ítems del test de comprensión verbal aplicado, para los valores θ_j estimados a los niños del grupo focal. ¿Funciona el ítem de modo diferencial para los niños urbanos y los rurales? Veamos cómo se procede para obtener el valor de *DP*.

θ_j	$P_R(\theta_j)$	$P_F(\theta_j)$	$[P_R(\theta_j) - P_F(\theta_j)]$
–1,0	0,25	0,10	0,15
–0,5	0,44	0,26	0,18
0,2	0,54	0,34	0,20
1,0	0,67	0,40	0,27
1,8	0,80	0,55	0,25
Total:			**1,05**

$$DP = \frac{1{,}05}{5} = 0{,}21$$

Este elevado valor de *DP* indica que el ítem funciona diferencialmente para los dos grupos, perjudicando claramente al grupo focal. Una vez detectado estadísticamente el FDI, la tarea sustantiva fundamental es averiguar las causas del FDI: ¿por qué este ítem perjudica tan claramente a los niños de origen rural? Nótese que las técnicas para detectar el FDI no informan sobre las posibles causas del FDI, y habrán de ser los expertos del campo quienes tengan que indagar esa cuestión.

Como ocurría en el método de las áreas, tampoco aquí existe una prueba estadística definitiva que nos diga cuándo el tamaño del índice *DP* es estadísticamente significativo. Por tanto, hay que seguir alguna de las estrategias citadas para el caso de las áreas.

El índice de diferencias de probabilidades (*DP*) puede modificarse ligeramente, elevando las diferencias al cuadrado; para expresarlo de modo que siempre sea positivo:

$$DP = \sum_{j=1}^{n_F} \frac{[P_R(\theta_j) - P_F(\theta_j)]^2}{n_F} \qquad [7.56]$$

Si se utiliza esta fórmula, hay que complementarla con la inspección gráfica de las CCI, pues solo indica el tamaño del FDI, pero no si es a favor del grupo de referencia o del focal, ni si el FDI es uniforme o no uniforme.

Nótese que ambos índices pueden aplicarse igualmente si se divide el rango de θ en varios intervalos, pudiendo promediarse su valor para cada uno de los intervalos y ofrecer así el tamaño de *DP* en función del rango de θ estudiado.

Según la autorizada opinión de Camilli y Shepard (1994), el índice *DP* sería el más indicado para el análisis del FDI bajo la óptica de la TRI, con lo cual es ciertamente difícil no estar de acuerdo.

Muestras pequeñas

Los métodos basados en la TRI tienen un inconveniente común a todos ellos, derivado de los propios modelos de TRI, y es que para estimar los parámetros con precisión las muestras tienen que ser bastante amplias, por encima de 500 personas, por dar un número aproximado. Esto no suele representar ningún problema para las grandes compañías que construyen y utilizan test, pero supone un serio inconveniente para numerosos profesionales, tales como psicólogos o educadores, no integrados en grandes organizaciones. Es muy frecuente que este tipo de profesionales disponga de pocas personas, sobre todo en el caso del grupo focal, con frecuencia minoritario, y, sin embargo, deseen indagar el FDI de las pruebas utilizadas. Suele recomendarse el uso de alguno de los métodos clásicos (Mantel-Haenszel, índice de estandarización, etc.) cuando se tienen muestras pequeñas, aunque tampoco estos funcionan a la perfección en esas circunstancias. El rendimiento de los modelos de TRI con muestras pequeñas no ha sido exhaustivamente investigado, aunque se dispone de varios trabajos que lo abordan (Hambleton et al., 1993; Linn y Harnisch, 1981; Mazor et al., 1992; Muñiz, Hambleton y Xing, 2001; Shepard et al., 1985).

Linn y Harnisch (1981) propusieron una variante interesante para seguir utilizando los modelos de TRI cuando en uno de los grupos, habitualmente el focal, el número de personas es reducido. Los pasos a seguir para la utilización de este procedimiento serían los siguientes:

1. Estimar los parámetros de los modelos y las puntuaciones θ de las personas a partir de la muestra completa, considerando conjuntamente el grupo de referencia y el focal.
2. Comparar las probabilidades $P(\theta)$ correspondientes a la CCI así estimada con el rendimiento real de las personas del grupo focal.

Si no existiese FDI, ambos valores deberían coincidir, excepto por variaciones debidas al azar. La discrepancia entre ambos valores es la medida del FDI.

El primer paso para obtener el índice de Linn y Harnisch es calcular las diferencias para cada persona del grupo focal (Z_j) entre su puntuación en el ítem (u_j), que tomará los valores de 1 o 0 según lo acierte o lo falle, y su probabilidad de superar el ítem según la CCI elaborada a partir del grupo total [$P_{F+R}(\theta_j)$]:

$$Z_j = \frac{[u_j - P_{F+R}(\theta_j)]}{\sqrt{P_{F+R}Q_{F+R}}} \qquad [7.57]$$

expresión que nos da la diferencia estandarizada para una persona *j* del grupo focal entre su puntuación empírica y la pronosticada por la CCI del grupo total. Promediando los valores Z_j de todas las personas del grupo focal, tendremos el índice general (Z_T) de FDI propuesto por Linn y Harnisch:

$$Z_T = \sum_{j=1}^{n_F} \frac{Z_j}{n_F} \qquad [7.58]$$

donde:

Z_j: Viene dado por la expresión 7.57.
n_F: Es el número de personas del grupo focal.
$\sum_{j=1}^{n_F}$: Va de 1 al número total de personas del grupo focal n_F.

Aparte de calcular este índice global de cada ítem, los autores sugieren que se calculen para zonas de interés especial dentro del rango de θ. Por ejemplo, podría dividirse θ en varios intervalos y calcular el índice Z para cada uno de ellos, lo que permitiría analizar el FDI en función de las distintas zonas de θ. Ponderando los valores medios de Z de cada intervalo por el número de personas del grupo focal dentro del intervalo correspondiente y sumando los productos, se reproduciría el valor global Z_T.

Si las personas del grupo focal obtienen sistemáticamente puntuaciones inferiores a las pronosticadas por la CCI global, el valor del índice Z_T será negativo; por el contrario, si su rendimiento es superior, el índice será positivo. En el caso de FDI no uniforme podrían anularse las diferencias posi-

tivas y las negativas, dando el índice una falsa idea de no FDI, por lo que de nuevo aquí se recomienda complementar el índice numérico con la representación gráfica.

Camilli y Shepard (1994) sugieren la posibilidad de elevar las diferencias Z_j al cuadrado para que el valor de Z_T sea siempre positivo, en cuyo caso la fórmula vendría dada por:

$$Z_T = \sqrt{\sum_{j=1}^{n_F} \frac{Z_j^2}{n_F}} \qquad [7.59]$$

Una discrepancia clara entre los resultados obtenidos mediante 7.58 y 7.59 sería indicativa de FDI no uniforme. De nuevo, aquí no se dispone de pruebas estadísticas adecuadas para determinar la significación estadística de ZT, por lo que ha de procederse de modo descriptivo, como ya se indicara en los casos anteriores.

Ejemplo

En la primera columna de la tabla adjunta aparecen las puntuaciones de las seis personas que componen el grupo focal. La segunda columna (u_j) refleja si acertaron (1) o fallaron (0) el ítem analizado. La tercera (P_{R+F}) ofrece la probabilidad de acertar el ítem según la CCI del grupo total.

Para calcular el índice de Linn y Harnisch lo primero que hay que hacer es obtener la diferencia entre u_j y P_{R+F} (cuarta columna) para cada persona y dividir estas diferencias entre $\sqrt{P_{F+R}Q_{F+R}}$ (quinta columna), obteniéndose así los valores Z_j.

Los valores de la última columna se calculan del siguiente modo:

$$Z_1 = \frac{-0{,}50}{\sqrt{(0{,}50)(0{,}50)}} = -1{,}00$$

$$Z_2 = \frac{0{,}36}{\sqrt{(0{,}64)(0{,}36)}} = 0{,}75$$

$$Z_3 = \frac{0{,}20}{\sqrt{(0{,}80)(0{,}20)}} = 0{,}50$$

$$Z_4 = \frac{0{,}13}{\sqrt{(0{,}87)(0{,}13)}} = 0{,}38$$

$$Z_5 = \frac{0{,}10}{\sqrt{(0{,}90)(0{,}10)}} = 0{,}33$$

$$Z_6 = \frac{0{,}08}{\sqrt{(0{,}92)(0{,}08)}} = 0{,}30$$

θ_j	u_j	P_{R+F}	$(u_j - P_{R+F})$	Z_j
–1,0	0	0,50	–0,50	–1,00
–0,5	1	0,64	0,36	0,75
0,2	1	0,80	0,20	0,50
1,0	1	0,87	0,13	0,38
1,8	1	0,90	0,10	0,33
2,3	1	0,92	0,08	0,30
Total:				**1,26**

A partir de los valores de la última columna, se obtiene el valor global del índice, dado por la fórmula 7.58:

$$Z_T = \frac{1{,}26}{6} = 0{,}21$$

Una diferencia media de 0,21 es bastante elevada, lo que indica que el ítem presenta un FDI para el grupo de referencia y el focal, a favor del focal, dado que el signo de Z_T es positivo. Nótese cómo los valores empíricos en el grupo focal son superiores, excepto en un caso, a los pronosticados por la CCI del grupo total.

La filosofía del procedimiento de Linn y Harnisch (1981) de juntar ambos grupos, referencia y focal, para llevar a cabo las estimaciones de los parámetros no está exenta de críticas. A nivel teórico puede ponerse en duda la legitimidad de juntar ambos grupos, si se sospecha de la existencia de FDI, pues no constituirían muestras aleatorias de la misma población. Desde el punto de vista aplicado, si el grupo de referencia es mucho más numeroso que el focal, es evidente que va a pesar más e influir en la estimación de los parámetros y, por tanto, va a tender a acercar la CCI global al grupo de referencia. No hay atajo sin trabajo, y esos serían los peajes a pagar para obtener estimaciones más estables que las que se conseguirían a partir de un grupo focal menguado en número de personas.

Comparación de los parámetros

Otro modo de evaluar el FDI es comparar los parámetros de la curva característica estimada para el ítem en cada subpoblación (Lord, 1980). El ítem estará insesgado en la medida en que los parámetros estimados en ambas subpoblaciones coinciden. Como señalan Hambleton y Swaminathan (1985, p. 293), la lógica de este procedimiento es equivalente a la del anterior, siendo atribuibles las posibles diferencias entre ambos a las distintas formas de operativizarlos.

Para el modelo de Rasch el único parámetro a comparar es b, y ello puede hacerse mediante el estadístico de contraste Z propuesto por Wright et al. (1976):

$$Z = \frac{\hat{b}_R - \hat{b}_F}{\sqrt{S^2(\hat{b}_R) + S^2(\hat{b}_F)}} \qquad [7.60]$$

donde $\hat{b}_R$ y $\hat{b}_F$ son los parámetros del ítem estimados en cada grupo, referencia y focal, $S^2(\hat{b}_R)$ y $S^2(\hat{b}_F)$ son las varianzas de $\hat{b}$ en cada grupo y Z se distribuye según la curva normal.

El valor de Z obtenido se compara con el de la distribución normal correspondiente al nivel de confianza adoptado, lo que permite aceptar o rechazar la hipótesis nula H_0: $b_1 = b_2$.

Para los modelos logísticos de *dos y tres parámetros* habrá que comparar b y a, considerando c invariante y con valor cero en el caso de dos parámetros. Lord (1980) propone el estadístico de contraste Z para ambos parámetros:

$$Z_b = \frac{\hat{b}_R - \hat{b}_F}{\sqrt{S^2(\hat{b}_R) + S^2(\hat{b}_F)}} \qquad [7.61]$$

$$Z_a = \frac{\hat{a}_R - \hat{a}_F}{\sqrt{S^2(\hat{a}_R) + S^2(\hat{a}_F)}} \qquad [7.62]$$

donde las varianzas de $\hat{a}$ y $\hat{b}$ para cada subpoblación vienen dadas por:

$$S^2_{\hat{b}} = \frac{1}{D^2\hat{a}^2/(1-\hat{c})^2} \sum_{j=1}^{N} \frac{[P_j(\theta) - \hat{c}]^2 Q_j(\theta)}{P_j(\theta)} \qquad [7.63]$$

$$S^2_{\hat{a}} = \frac{1}{D^2/(1-\hat{c})^2} \sum_{j=1}^{N} \frac{(\hat{\theta}_j - \hat{b})^2 [P_j(\theta) - \hat{c}]^2 Q_j(\theta)}{P_j(\theta)} \qquad [7.64]$$

y donde a, b, θ, D, $P_j(\theta)$, $Q_j(\theta)$ son los bien conocidos valores de los modelos logísticos, N es el número de personas que alcanzan el ítem y θ_j es el valor estimado de θ para cada persona.

Ahora bien, más interesante que comparar cada parámetro por separado es hacerlo conjuntamente valiéndose de χ^2, en cuyo caso la hipótesis nula será: $b_R = b_F$ y $a_R = a_F$.

Según Lord (1980):

$$\chi^2 = V\Sigma^{-1}V' \qquad [7.65]$$

donde:

- χ^2: Tiene dos grados de libertad.
- V: Vector de dimensión (1 × 2) de las diferencias entre los parámetros b y a de ambas subpoblaciones.
- V': Vector traspuesto de V.
- Σ^{-1}: Inversa de la matriz suma de varianzas covarianzas de V para ambos grupos, de dimensión (2 × 2).

Si la fórmula de Lord se aplica al modelo logístico de un parámetro, se simplifica notablemente, pudiendo expresarse:

$$\chi^2 = \frac{(b_F - b_R)^2}{\text{var}(b_F) + \text{var}(b_R)} \qquad [7.66]$$

donde b_F y b_R son las estimaciones de los parámetros b en los dos grupos, y var (b_F) y var (b_R), las varianzas estimadas, dadas por la inversa de las funciones de información correspondiente a los parámetros de dificultad estimados.

EJEMPLO

Los parámetros de un ítem para dos grupos, estimados mediante el modelo logístico de dos parámetros, fueron los siguientes:

$$a_R = 0{,}40 \quad b_R = 0{,}42$$
$$a_F = 0{,}90 \quad b_F = 1{,}42$$

En el grupo de referencia las varianzas y covarianzas de las estimaciones de los parámetros fueron: var (a) = 0,02; var (b) = 0,01; cov (a, b) = 0,03. En el grupo focal se obtuvieron los siguientes valores: var (a) = 0,05; var (b) = 0,07; cov (a, b) = 0,03. Veamos cómo se procede para obtener el valor de χ^2.

El vector V vendría dado por las diferencias entre los parámetros a y b:

$$[(0{,}90 - 0{,}40), (1{,}42 - 0{,}42)] = [0{,}50, 1{,}00]$$

El vector V' es el transpuesto de V:

$$\begin{bmatrix} 0{,}50 \\ 1{,}00 \end{bmatrix}$$

Para obtener la matriz Σ^{-1} han de sumarse las matrices de varianzas covarianzas en el grupo de referencia y el focal y posteriormente calcular la inversa de la suma. La matriz de varianzas covarianzas para el grupo de referencia según los datos del enunciado vendría dada por:

$$\begin{bmatrix} 0{,}02 & 0{,}03 \\ 0{,}03 & 0{,}01 \end{bmatrix}$$

Para el grupo focal:

$$\begin{bmatrix} 0{,}05 & 0{,}03 \\ 0{,}03 & 0{,}07 \end{bmatrix}$$

La suma de las matrices anteriores nos da la matriz Σ:

$$\begin{bmatrix} 0{,}07 & 0{,}06 \\ 0{,}06 & 0{,}08 \end{bmatrix}$$

Se calcula la inversa de la matriz Σ:

— Determinante:

$$(0{,}07)(0{,}08) - (0{,}06)(0{,}06) = 0{,}002$$

— Inversa:

$$\begin{bmatrix} 0{,}08/0{,}002 & -0{,}06/0{,}002 \\ -0{,}06/0{,}002 & 0{,}07/0{,}002 \end{bmatrix}$$

Efectuando las divisiones, se obtiene finalmente la matriz Σ^{-1}:

$$\begin{bmatrix} 40 & -30 \\ -30 & 35 \end{bmatrix}$$

Ahora ya disponemos de los datos necesarios para obtener χ^2, que vendrá dada por el producto del vector V, la matriz Σ^{-1} y el vector V':

$$\chi^2 = V\Sigma^{-1}V' = [0{,}50, 1{,}00] \times \begin{bmatrix} 40 & -30 \\ -30 & 35 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0{,}50 \\ 1{,}00 \end{bmatrix} = 15$$

Con dos grados de libertad el valor de χ^2 en las tablas correspondientes al nivel de confianza de 95% viene dado por 5,99. Dado que nuestro valor empírico (15) es muy superior, rechazamos la hipótesis nula de que los parámetros del ítem son estadísticamente iguales para ambos grupos, es decir, el ítem presenta un funcionamiento diferencial.

Mediante χ^2 puede someterse a prueba la hipótesis no solo de que $b_1 = b_2$ y $a_1 = a_2$, sino que también $c_1 = c_2$, en cuyo caso los grados de libertad de χ^2 serían tres y no dos. La razón por la que en la práctica no suele incluirse c es debido a que su estimación es bastante imprecisa, tiene un error típico alto, por lo que su inclusión aumentaría el conservadurismo de la prueba, rebajando la posibilidad de detectar ítems con FDI. Si realmente existe un FDI, ha de manifestarse en los valores de a y b, y si estos no fuesen distintos, sería demasiado arriesgado afirmar la existencia de FDI basándonos únicamente en el parámetro c.

La crítica fundamental a estas técnicas de comparación de parámetros es que pueden detectar FDI cuando, en realidad, las diferencias entre las CCI son mínimas en el rango de las puntuaciones θ de interés; además, el número de falsos positivos es elevado, detecta muchos ítems con FDI que no son tales. Por ello, estas técnicas de comparación de los parámetros de los ítems no son las más utilizadas en la práctica.

Reflexiones finales

Vistos sumariamente el concepto y algunos de los procedimientos más sencillos para la evaluación del FDI, no se le habrá escapado al lector la contradicción de fondo en la que parece moverse la TRI al tratar de captar el FDI. Los ítems contenidos en un banco estarán todos calibrados en la misma métrica, y si se han incluido será porque se ajustan al modelo, entre otras serán unidimensionales para la población en la que se calibraron. Se supone que en dicha población se han incluido las subpoblaciones de interés en las que se desea estudiar el funcionamiento diferencial. Si ello es así, *teóricamente* nunca se podrían encontrar curvas características distintas para un ítem al emplear en la estimación muestras distintas, se violaría el principio de invarianza de los parámetros y de unidimensionalidad. Es decir, si un ítem se ha calibrado como unidimensional para una población de personas, es una incongruencia lógico-teórica admitir la posibilidad de curvas características distintas para ese ítem cuando se utilizan dos subpoblaciones en la estimación. Se violan claramente los presupuestos de los modelos de TRI. De hecho, el propio concepto de FDI encaja mal con la formulación teórica de la TRI para modelos unidimensionales. La mayoría de autores apenas se plantean esta cuestión de carácter teórico, con notoria excepción de Lord (1980), que propone cinco consejos para mitigarla empíricamente en lo que él denomina «purificación del test», terminando su exposición en tono escéptico al señalar que ni las pruebas estadísticas son óptimas, ni válida la estimación de los parámetros de los ítems sesgados. Claro que desde que Lord dijera esas palabras, la tecnología para la detección de FDI ha avanzado considerablemente. Aquí se han expuesto algunos métodos sencillos para ejemplificar la lógica de la evaluación de FDI, pero en la actualidad disponemos de métodos con sus correspondientes programas informáticos mucho más sofisticados. Aunque el mejor de los métodos de detección del FDI es prevenirlo, una buena guía para tratar de identificar los ítems con funcionamiento diferencial puede verse en Hambleton (2006) o Sireci y Rios (2013). Por su parte, Rogers y Swaminathan (2016), buenos conocedores del campo, han hecho una excelente síntesis conceptual de las distintas etapas por las que ha pasado el estudio del FDI, apuntando algunas reflexiones de futuro. También Zumbo (2007b) analiza el pasado, presente y futuro del FDI. Buenos tratamientos pueden verse en textos como los de Berk (1982), Holland y Wainer (1993), Camilli y Shepard (1994), Camilli (2006) y Osterlind y Everson (2009). En español véanse Fidalgo (1996) y Gómez, Hidalgo y Guilera (2010), donde se puede encontrar una relación de los programas informáticos más habituales para la detección del FDI. El cuadro que se presenta a continuación da una buena panorámica del estado actual de las investigaciones sobre el FDI:

Clasificación de los métodos de funcionamiento diferencial de los ítems según Rogers y Swaminathan (2016)

		Variable de respuesta	
Variable de contraste		**Discreta**	**Continua**
Empírica	**Discreta**	Mantel-Haenszel SIBTEST, STD	SIBTEST
	Continua	Regresión logística	Regresión logística, Path Analysis
Latente	**Discreta**	Modelos de rasgo latente	Análisis de perfiles latentes
	Continua	TRI, modelos de ecuaciones estructurales	Modelos de ecuaciones estructurales

Como ocurre con cualquier otra metodología, o tecnología estadística, los métodos para detectar el FDI hay que utilizarlos con sentido común y prudencia, pues no constituyen una varita mágica que nos da automáticamente la solución; así Sireci y Rios (2013) nos dan unos sabios consejos para la detección del FDI:

1. Hay que seleccionar el método para detectar el FDI que mejor encaje en el tipo de datos que se maneja, pues hay características de los datos que afectan a los métodos utilizados, tales como el número de catego-

rías de los ítems, el tamaño de la muestra o la dimensionalidad de los datos.
2. Cuando sea posible utilizar más de un método de detección para confirmar los resultados.
3. Utilizar algún indicador del tamaño del efecto para distinguir entre la mera significación estadística y un efecto relevante.
4. Utilizar representaciones gráficas para interpretar y comunicar adecuadamente el FDI.
5. Cuando se encuentren ítems con FDI, eliminarlos del test (purificar el test) y repetir los análisis, llevando a cabo el proceso de forma gradual o iterativa.
6. Utilizar muestreos adecuados que permitan confirmar los resultados obtenidos.
7. Extraer varias muestras aleatorias del grupo de referencia cuando las diferencias en la variable medida entre el grupo de referencia y el focal sean importantes; por ejemplo los valores centrales difieren en más de una desviación típica.
8. A la hora de interpretar los resultados del FDI hay que tener en cuenta la dirección del FDI, los posibles efectos de amplificación o cancelación, así como sus posibles implicaciones a la hora de interpretar las puntuaciones de la prueba.

EJERCICIOS

1. En la tabla adjunta aparecen los parámetros *a, b* y *c* de seis ítems estimados mediante el programa BILOG para un grupo de mujeres (referencia) y otro de hombres (focal). Los seis ítems corresponden a un test de comprensión verbal.

Ítems	Referencia			Focal		
	a	*b*	*c*	*a*	*b*	*c*
1	0,46	-0,79	0,11	0,46	-0,79	0,11
2	0,66	0,60	0,30	0,66	0,60	0,30
3	0,49	0,02	0,17	0,49	1,02	0,17
4	0,91	-0,44	0,26	0,91	-0,44	0,26
5	1,15	0,38	0,18	1,15	0,38	0,18
6	0,40	0,52	0,22	0,90	1,52	0,22

1. Mediante una inspección visual de la tabla, ¿considera que alguno de los ítems parece mostrar un funcionamiento diferencial para ambos grupos? Razone su respuesta.
2. Represente gráficamente las curvas características de los seis ítems. Utilice para ello la fórmula del modelo logístico de tres parámetros.
3. Calcule mediante la fórmula de Raju el área comprendida entre las CCI de los seis ítems para los grupos de referencia y focal. ¿Qué ítems muestran FDI?
4. Calcule el funcionamiento diferencial de los seis ítems mediante χ^2 de Lord y compare los resultados con los obtenidos por el método de Raju en el apartado anterior. En el grupo de referencia las varianzas de las estimaciones (para todos los ítems) son: $\text{var}(a) = 0{,}02$, $\text{var}(b) = 0{,}01$ y $\text{cov}(a,b) = 0{,}03$. En el caso del grupo focal (para todos los ítems), $\text{var}(a) = 0{,}05$, $\text{var}(b) = 0{,}07$, $\text{cov}(a,b) = 0{,}03$.
5. Calcule la correlación entre los parámetros *b* de los ítems de ambos grupos. Elimine los ítems que muestren FDI y vuelva a calcular la correlación. Explique los cambios, si es que los hubiere, experimentados por el valor de la correlación.

2. Las cuatro personas de un grupo focal obtuvieron en un ítem las siguientes puntuaciones: 0, 1, 1, l. Es decir, solo la primera lo falló. Las puntuaciones de estas personas en el test fueron respectivamente: –1, 0, 0,5, 1,0. Calcule el índice propuesto por Linn y Harnisch para el citado ítem, sabiendo que sus parámetros estimados en el grupo total fueron: $a = 0{,}6$, $b = 1$, $c = 0{,}2$.

SOLUCIONES

1.1. Ítems 3 y 6.
3. 0, 0, 0,83, 0, 0, 1,11.
4. 0, 0, 35, 0, 0, 15.
5. 0,83, 1.

2. 0,59.

11. TEST ADAPTATIVOS INFORMATIZADOS

11.1. Concepto

Otra de las aplicaciones directas de la TRI ha sido la potenciación de un tipo de test que, si bien eran conocidos y utilizados de antiguo, encontraban serias dificultades dentro del marco de la psicometría clásica, los *test adaptativos* o *test a medida.* Estos test suelen encontrarse también citados por otros nombres, tales como «test adaptados a la persona», «test de nivel flexible», «test ramificados», «test individualizados», «test programados» o «test secuenciales», pero su característica más definitoria es que se construyen a la medida, adaptados al tipo de personas a las que se aplica. En la utilización habitual y clásica de los test, se aplica la misma prueba a todas las personas, independientemente de su nivel en la variable medida. Ahora bien, esto no es lo más deseable, pues las mediciones serán más eficientes si el nivel de dificultad del test se adapta al nivel de competencia de la persona porque se evita, en gran parte, la tendencia de las personas a contestar al azar y desmotivarse cuando los ítems exceden notablemente sus conocimientos, así como al aburrimiento si los ítems son demasiado fáciles.

Los test adaptativos son unos test que, aun midiendo la misma variable, no son idénticos para todas las personas, varían en función del nivel de competencia que estén destinados a medir.

Nótese que eso no es nada nuevo conceptualmente en psicología, todo lo contrario. El pionero test de Binet, por ejemplo, es en cierto modo de este tipo: la secuencia de tareas que se presenta al niño depende de sus respuestas previas. Los métodos psicofísicos clásicos constituyen otro ejemplo de adaptación a la persona a la hora de estimar los umbrales (Muñiz, 1991).

Lo que aportará la TRI de nuevo a la medición adaptativa será la posibilidad de equiparar convenientemente puntuaciones obtenidas con distintos instrumentos para la misma variable, así como pautas adecuadas para ajustar el test a la persona. Adicionalmente, los nuevos tiempos también aportan la posibilidad de hacerlo todo en un ordenador. Ello ha provocado que a estos test adaptativos se les denomine «test adaptativos informatizados» (TAI), lo cual puede conducir a error, pues el que estén informatizados no es una característica definitoria central, su esencia es la adaptación de la prueba a la persona evaluada. La informatización de un test, esto es, su aplicación y análisis mediante ordenador, no implica que sea adaptado a las personas; de hecho, en la actualidad la mayoría de los test clásicos disponen de una versión informatizada. Véanse, por ejemplo, los catálogos de las editoriales españolas de test, tales como Tea, Pearson, Cepe o Eos.

11.2. Desarrollo

Un test adaptativo, como ya se ha apuntado, conlleva dos tareas fundamentales: en primer lugar, su construcción, es decir, decidir qué ítems lo integran, y, en segundo lugar, expresar los resultados en una métrica común a otros test diferentes utilizados con otras personas para medir la misma variable. La TRI constituye el marco adecuado para resolver ambos, pues al disponer de un banco de ítems calibrados, es decir, con parámetros conocidos, permite elegir los oportunos según el nivel θ de la persona a evaluar. Utilizando la función de información, se elegirán aquellos ítems que proporcionen más información para el nivel θ de la persona; además, los resultados vendrán expresados en la misma escala (la métrica del banco), independientemente del test empleado. Ahora bien, para adaptar el test a la persona con una determinada θ, el problema consiste en estimar aproximadamente el nivel θ de la persona, que obviamente no se conoce a priori, y que es

lo que pretendemos evaluar. En suma, ¿cómo se adapta una prueba al nivel de la persona a evaluar si no conocemos ese nivel? Existe toda una literatura especializada al respecto; véanse, por ejemplo, Wainer (1990), Lord (1980), Urry (1977), Weiss (1983), y en español Olea, Ponsoda y Prieto (1999). En términos generales, pueden distinguirse dos aproximaciones principales:

— Doble-nivel.
— Multinivel.

En la lógica del *doble-nivel,* en un primer paso, estadio o nivel se aplica el mismo test a todas las personas y, en función de la puntuación obtenida, se les aplica un segundo test según haya sido su rendimiento en el primero. Es decir, el primer test común se utiliza para estimar de forma aproximada el nivel θ de las personas, que se medirá con precisión mediante la segunda aplicación. Esta estrategia tiene su lógica, pero en la actualidad prácticamente no se utiliza.

En la estrategia de *multinivel,* aunque caben muchas alternativas, en general se va aplicando ítem a ítem, decidiendo cuál será el siguiente en función de las respuestas a los anteriores. Los aciertos conducirán a ítems más difíciles, y los fallos, a ítems más fáciles. Ello plantea toda una tecnología y decisiones a tomar, tales como cuándo detener el proceso, cómo ramificar los caminos, cuántos ítems aplicar, etc., muy bien tratado por Lord (1980) y en español Olea et al. (1999).

Además de la ventaja central de los test adaptados, consistente, como ya se ha señalado, en el uso de test ajustados al nivel de la persona para maximizar así la información a ese nivel, otros beneficios colaterales no son despreciables. Normalmente, se necesita *menor número de ítems,* lo que disminuye los efectos de la fatiga. *Aumenta la motivación* de las personas, pues ni los muy competentes han de contestar ítems demasiado fáciles para ellos ni los menos competentes se estrellarán ante ítems excesivamente difíciles para su nivel.

Ejemplo

Vamos a ilustrar con un ejemplo la lógica descrita.

A continuación se presenta una tabla en la que aparecen los parámetros *a*, *b* y *c* estimados a los cuarenta ítems de un banco de comprensión verbal. También aparecen en la tabla el índice de dificultad clásico (*p*) y el índice de discriminación clásico o correlación ítem-test (r_{jx}). Estos datos, aparte de otros, siempre acompañan a los ítems en un banco.

Ítem	*a*	*b*	*c*	*p*	r_{jx}
1	0,46	–0,79	0,12	0,66	0,44
2	0,51	–0,59	0,12	0,64	0,46
3	0,47	–1,07	0,12	0,70	0,44
4	0,65	–1,35	0,12	0,78	0,54
5	0,67	–1,57	0,12	0,82	0,56
6	0,67	–0,33	0,18	0,62	0,50
7	0,39	0,67	0,00	0,37	0,41
8	0,63	–0,07	0,17	0,57	0,47
9	0,67	0,61	0,30	0,53	0,39
10	0,42	–1,22	0,12	0,71	0,44
11	0,51	–0,78	0,12	0,66	0,47
12	1,10	0,37	0,18	0,44	0,56
13	0,84	–0,29	0,20	0,62	0,57
14	1,17	0,06	0,25	0,56	0,56
15	0,93	–0,24	0,21	0,62	0,57
16	0,74	–0,01	0,09	0,51	0,56
17	1,01	–0,13	0,26	0,63	0,57
18	0,95	0,96	0,24	0,40	0,43
19	0,96	0,09	0,19	0,53	0,57
20	1,07	0,19	0,19	0,51	0,56
21	0,60	–0,43	0,12	0,61	0,49
22	0,68	0,94	0,17	0,38	0,44
23	0,80	0,12	0,12	0,50	0,54
24	1,11	1,64	0,15	0,22	0,35
25	0,83	0,97	0,24	0,40	0,42
26	0,92	0,98	0,14	0,32	0,48
27	0,79	0,61	0,21	0,45	0,46
28	0,57	–1,18	0,14	0,74	0,49
29	0,65	–0,67	0,14	0,66	0,53
30	0,48	–0,79	0,14	0,66	0,45
31	0,61	–0,78	0,14	0,69	0,49
32	0,61	–0,20	0,14	0,58	0,53
33	0,82	–0,25	0,27	0,66	0,54
34	0,57	–0,10	0,14	0,55	0,47
35	0,52	–0,62	0,14	0,65	0,46
36	0,91	–0,45	0,26	0,70	0,58
37	0,87	0,63	0,37	0,54	0,44
38	0,63	0,38	0,21	0,51	0,46
39	0,69	–1,30	0,14	0,79	0,57
40	0,73	0,09	0,18	0,54	0,53

Supongamos tres personas (X, Y, Z) a las que a partir de datos previos se estima tentativamente que les corresponderían las siguientes puntuaciones respectivamente: $X = –1{,}3$, $Y = 0{,}0$, $Z = 0{,}9$. Si hubiese que elegir únicamente cinco ítems para evaluar a cada una de estas tres personas, ¿cuáles se elegirían? En otras palabras, ¿cuál sería el test adaptativo de cinco ítems que evaluaría con mayor precisión a cada una de esas tres personas? Como se ha señalado, se trata de elegir aquellos ítems cuya dificultad sea más parecida a las puntuaciones θ de las personas. La razón es bien sencilla, pues, como se expuso al tratar de la función de información, en los modelos de uno y dos parámetros los ítems miden con mayor precisión, dan su información máxima, exactamente para el valor de $\theta = b$; en el caso del modelo de tres parámetros, hay que corregir ese valor ligeramente.

Compruebe el lector cómo para evaluar a la persona X, cuya puntuación $\theta = –1{,}3$, los cinco ítems más adecuados serían el 4, 5, 10, 28 y 39. Se trataría de los ítems cuya dificultad (b) se acerca más a la competencia de la persona ($\theta = –1{,}3$). Sin embargo, el test que mejor se adapta a la persona Y ($\theta = 0{,}0$) estaría formado por los ítems: 8, 14, 16, 19 y 40. Finalmente, para evaluar a Z se elegirían los ítems 7, 18, 22, 25 y 26. Nótese que ninguno de los ítems es compartido por los test elegidos para cada persona, lo cual es debido a que el nivel de las tres personas en la variable medida θ es muy distinto, por lo que los ítems elegidos también lo son.

A pesar de usar tres test distintos, los resultados vendrían expresados en una métrica común, pues los tres se han extraído de un banco cuya métrica es única para todos los ítems que contiene. En la práctica no se procede habitualmente como se ha hecho aquí con fines ilustrativos, sino que se van presentando (mediante un ordenador) los ítems uno a uno y, según la persona vaya acertando o fallando, se aumenta o disminuye la dificultad del ítem siguiente.

A cada paso el programa de ordenador estima la puntuación θ de la persona y la precisión de la estimación, deteniéndose el proceso cuando se alcanza una precisión prefijada mediante una función de información objetivo, aunque también se pueden establecer otros criterios. En la práctica suele fijarse un número mínimo de ítems que toda persona debe contestar, aunque se haya alcanzado con menos ítems la precisión prefijada. Se hace para mejorar la validez aparente de la prueba, pues las personas examinadas, ajenas a los arcanos psicométricos, no entienden bien que con tan pocos ítems puedan ser evaluadas con rigor.

En la actualidad los TAI constituyen una de las líneas de trabajo más vigorosas dentro de la evaluación psicométrica, tanto en investigación como en las aplicaciones profesionales. Se han desarrollado TAI para prácticamente todos los campos aplicados dentro de la psicología, educación y en general ciencias sociales y de la salud. Por ejemplo, se han propuesto TAI para evaluar aspectos del cáncer (Petersen et al., 2006), pediatría (Allen, Ni y Haley, 2008), dolor de espalda (Kopec et al., 2008), ansiedad (Gibbons et al., 2008; Walter et al., 2007), depresión (Smits, Cuijpers y Van Straten, 2011), esquizotipia (Fonseca, Menéndez, Paino, Lemos y Muñiz, 2013), calidad de vida (Rebollo et al., 2009, 2010), satisfacción laboral (Chien et al., 2009) o clima organizacional (Menéndez, Peña, Fonseca y Muñiz, 2017), solo por citar algunos ejemplos. El aumento del uso de los TAI a partir de los años ochenta ha sido exponencial, con millones de aplicaciones y numerosas pruebas nacionales e internacionales de todo tipo aplicadas en forma de TAI (Wainer, 2000; Zenisky y Luecht, 2016).

Para quienes deseen profundizar en el estudio de los TAI hay una literatura abundante. En español el libro editado por Olea, Ponsoda y Prieto (1999) es de lectura obligada, y en él se abordan los aspectos fundamentales de los TAI, incluido el *software* disponible para implementarlos. Una buena panorámica, también en español, puede verse en Barrada (2012). Una excelente visión del desarrollo histórico de los TAI puede consultarse en Way y Robin (2016), y sobre la situación actual y las perspectivas de futuro son de gran interés las reflexiones de Mills y Breithaupt (2016) y Zenisky y Luecht (2016). Abundan los textos clásicos para profundizar en el tema, entre los que cabe citar, por ejemplo, los de Bartram y Hambleton (2006), Davey (2011), Van der Linden y Glas (2010) o Yan, Von Davier y Lewis (2014). Cabe señalar, finalmente, que se está abriendo paso con fuerza un tipo de test adaptativos que no toman como unidad adaptativa los ítems individuales, sino grupos de ítems, por lo que se denominan «test multietápicos». Una buena exposición sobre ellos puede verse, por ejemplo, en Zenisky, Hambleton y Luecht (2010) y Yan et al. (2014).

EJERCICIOS

1. En la tabla adjunta aparecen los parámetros de los ítems correspondientes a un banco de ítems:

Ítem	*a*	*b*	*c*
1	0,46	–0,79	0,12
2	0,51	–0,59	0,12
3	0,47	1,07	0,12
4	0,65	–1,35	0,12
5	0,67	–1,57	0,12
6	0,67	0,33	0,18
7	0,39	0,67	0,00
8	0,63	–0,07	0,17
9	0,67	0,61	0,30
10	0,42	–1,22	0,12
11	0,51	–0,78	0,12
12	1,10	0,37	0,18
13	0,84	–0,29	0,20
14	1,17	0,06	0,25
15	0,93	–0,24	0,21
16	0,74	–0,01	0,09
17	1,01	–0,13	0,26
18	0,95	0,96	0,24
19	0,96	0,09	0,19
20	1,07	0,19	0,19
21	0,60	–0,43	0,12
22	0,68	0,94	0,17
23	0,80	0,12	0,12
24	1,11	1,64	0,15
25	0,83	0,97	0,24
26	0,92	0,98	0,14
27	0,79	0,61	0,21
28	0,57	–1,18	0,14
29	0,65	–0,67	0,14
30	0,48	–0,81	0,14
31	0,61	–0,78	0,14
32	0,61	–0,20	0,14
33	0,82	–0,25	0,27
34	0,57	–0,10	0,14
35	0,52	–0,62	0,14
36	0,91	–0,45	0,26
37	0,87	0,63	0,37
38	0,63	0,38	0,21
39	0,69	–1,30	0,14
40	0,73	0,09	0,18

1. A una persona se le ha estimado tentativamente mediante un pretest una puntuación $\theta = -1{,}22$. Imagínese por unos momentos que es usted un ordenador que tiene que aplicar un test adaptativo a esa persona. El ordenador se ha programado, como ahora usted, para que vaya presentando los ítems más adecuados para evaluar a esa persona. Tiene órdenes de detener el proceso, usted también, cuando la precisión alcanzada, expresada por la función de información del test, alcance el valor de 0,75. Indique qué ítems (y por qué orden) aplicaría a esa persona en dos supuestos: 1) que acierta todos los ítems que usted le va presentando, 2) que acierta uno sí y otro no, empezando por sí. (Exactamente la programación del ordenador es bastante más compleja, pues a cada paso reestima la puntuación de la persona evaluada, pero a usted, solidaridad humana, se le permite mantener fija θ en –1,22, a efectos de poder calcular la información aportada por cada ítem presentado.)
2. Calcule el error típico del test utilizado para evaluar a esa persona en las dos situaciones descritas.
3. Si se desea que el error típico del test sea 0,5, ¿cuál tiene que ser la cantidad de información aportada por el test?

SOLUCIONES

1.1. 1: 10, 28, 3, 30, 1, 31.
2: 10, 28, 39, 3, 4.

2. 1,12, 1,05.
3. 4.

8 Fases para la construcción de un test

Una vez que se han visto en los capítulos precedentes las propiedades psicométricas más importantes de los test, vamos a recapitular y presentar aquí de forma sintética los pasos generales que habría que seguir para construir un instrumento de medida. No se trata de una exposición exhaustiva, que excede las pretensiones de este texto introductorio, pero esperamos que permita al lector extraer una idea cabal de cómo proceder si tuviese que desarrollar un nuevo test, escala o cuestionario. Tratamientos exhaustivos pueden verse en los trabajos de Downing (2006), Downing y Haladyna (2006), Haladyna y Rodríguez (2013), Schmeiser y Welch, (2006) o Lane, Raymond y Haladyna (2016), entre otros muchos. En este apartado seguiremos en líneas generales los trabajos previos sobre el tema de Muñiz y Fonseca (2008, 2017).

La construcción de un instrumento de medida es un proceso complejo que aquí vamos a concretar en diez pasos, si bien estos no son automáticos ni universales y pueden variar en función del propósito del instrumento de medida (selección, diagnóstico, etc.), del modelo psicométrico utilizado (teoría clásica, teoría de respuesta a los ítems), del tipo de respuesta exigida por los ítems (selección o construcción), del formato de administración (lápiz y papel o informatizado) o del contexto de evaluación (diagnóstico, evaluación de rendimientos, etc.), por citar solo algunos casos. Todo el proceso de construcción debe desarrollarse de forma rigurosa y objetiva, siguiendo unos estándares de calidad, para así maximizar la validez de las inferencias hechas a partir de las puntuaciones obtenidas en la prueba por las personas evaluadas (Downing, 2006; Lane, Raymond y Haladyna, 2016). Puede decirse que el proceso de validación ya comienza a fraguarse incluso antes de la propia elaboración empírica del instrumento, pues todas las acciones que se realicen antes, durante y después permitirán recoger evidencias que ayuden a la interpretación de las puntuaciones y a la posterior toma de decisiones (Elosua, 2003; Markus y Borsboom, 2013; Muñiz, 2004; Wells y Faulkner-Bond, 2016; Zumbo, 2007a).

A continuación se sintetiza en diez pasos el procedimiento a seguir para desarrollar una prueba, que en esencia recogen las recomendaciones de los últimos estándares de la AERA, APA y NCME (2014). Autores como Downing (2006) y Lane et al. (2016) prefieren establecer doce pasos o fases. Por supuesto no existe un número mágico al respecto, lo esencial queda recogido en los diez propuestos. En la tabla 8.1 se recogen de forma esquemática los pasos que se deben considerar en el proceso de construcción y validación de un instrumento de medida.

A continuación se comentan brevemente los pasos propuestos en la tabla 8.1.

1. MARCO GENERAL

Todo proceso de construcción de un instrumento de medida comienza por una explicación detallada y precisa de cuáles son las razones que motivan su desarrollo. Un nuevo instrumento no se construye porque sí, hay que justificarlo adecuadamente. Asimismo, hay que delimitar con claridad cuál es la variable objeto de medición, cuál va a ser el contexto de aplicación, las circunstancias en las que se va a administrar el instrumento de evaluación, el tipo

TABLA 8.1

Fases del proceso de construcción de un test

1. Marco general

— Justificación y motivación.
— Contexto de aplicación.
— Uso e interpretación de las puntuaciones.

2. Definición de la variable medida

— Definición operativa.
— Definición sintáctica.
— Definición semántica.

3. Especificaciones

— Requerimientos de aplicación.
— Tipo, número, longitud, formato, contenido y distribución de los ítems.
— Especificaciones e instrucciones en la entrega del material, seguridad.

4. Construcción de los ítems

— Principios generales para la construcción de ítems.
— Tipos de ítems.
— Directrices para la construcción de ítems de elección múltiple.

5. Edición

— Composición.
— Edición.
— Puntuación y corrección.

6. Estudios piloto

— Selección de la muestra piloto (cualitativo y cuantitativo).
— Análisis y resultados del estudio piloto (cualitativo y cuantitativo).
— Depuración, revisión, modificación o construcción de ítems.

7. Selección de otros instrumentos de medida

— Justificación teórica.
— Obtener evidencias de relación con variables externas.
— Utilizar pruebas ya validadas.

8. Aplicación del test

— Selección y tamaño de la muestra y tipo de muestreo.
— Aplicación del instrumento de medida.
— Control de calidad y seguridad de la base de datos.

9. Propiedades psicométricas

— Análisis de los ítems.
— Fiabilidad.
— Validez.

10. Versión final del test

— Informe.
— Prueba final.
— Manual.

de administración (individual, colectiva), el formato de aplicación (lápiz y papel, informática) y qué decisiones se van a tomar a partir de las puntuaciones. Las causas que pueden llevar a la construcción de un instrumento de evaluación son lógicamente diversas. Por ejemplo, un psicólogo puede decidir construir un instrumento porque no existe ningún otro para medir una determinada variable, porque los instrumentos existentes presentan unas pésimas propiedades psicométricas, porque no incorporan alguna faceta relevante para analizar dicha variable o simplemente porque los existentes se han quedado obsoletos. Wilson (2005) detalla y comenta las principales razones para generar nuevos instrumentos de medida.

Los responsables de la construcción del instrumento de medida no solo deben especificar el motivo por el cual quieren desarrollar una nueva herramienta de medida sino que también deben delimitar con claridad cuál es el contexto en el que se va a aplicar, lo que incluye necesariamente la población objeto de medición (pacientes, alumnos, empresas, departamentos, etc.) y las circunstancias de aplicación (lugar, medios de los que se dispone y condiciones de aplicación, individual o colectiva). También debe especificarse de antemano con qué propósito van a ser utilizadas las puntuaciones y qué decisiones se van a tomar a partir de ellas. En este sentido, las puntuaciones en un instrumento de evaluación pueden servir para propósitos varios, como por ejemplo: seleccionar, diagnosticar, clasificar, orientar, evaluar un dominio específico o incluso como método de cribado (AERA, APA y NCME, 2014). Se debe dejar claro que las inferencias que se extraigan de las puntuaciones de un instrumento de medida no son universales, son siempre para un uso, contexto y población determinados. Nótese que lo que puede ser válido para un grupo determinado de personas o población tal vez no lo sea para otra, y lo que pueda ser válido en un contexto de evaluación no tiene por qué serlo en otro diferente (Zumbo, 2007).

En suma, un instrumento de medida vale para lo que vale, y hay que explicitarlo de forma clara. Ello no es óbice para que una prueba desarrollada originalmente con una determinada finalidad se revele en el futuro, tras distintos procesos de validación, como buena predictora de otros aspectos inicialmente no contemplados. Los usos que se hagan de una prueba deben venir avalados por evidencias empíricas, como bien establece la norma ISO 10667, relativa a la evaluación de personas en entornos laborales y organizacionales. Más aún, como indica nuestro código deontológico en su artículo 17, el psicólogo tiene que estar profesionalmente preparado y especializado en la utilización de métodos, instrumentos, técnicas y procedimientos que adopte en su trabajo, y debe reconocer los límites de su competencia y los de sus técnicas.

2. DEFINICIÓN DE LA VARIABLE MEDIDA

El objetivo esencial de esta segunda fase es la definición operativa, semántica y sintáctica de la variable medida, así como las facetas o dimensiones que la componen (AERA, APA y NCME, 2014; Carretero y Pérez, 2005; Wilson, 2005).

La variable evaluada debe definirse en términos operativos para que pueda ser medida de forma empírica (Muñiz, 2004). En este sentido, tan interesante puede ser definir cuidadosamente lo que es como lo que no es. La facilidad o dificultad de la definición operativa depende en cierta medida de la naturaleza de la variable objeto de medición. Para llevar a cabo una definición operativa es clave realizar una revisión exhaustiva de la literatura publicada al respecto, así como la consulta a expertos (Clark y Watson, 1995; Wilson, 2005). Ello permite, por un lado, delimitar la variable objeto de medición, y considerar todas sus dimensiones relevantes, y, por otro, identificar con claridad los comportamientos más representativos de tal variable (Calero y Padilla, 2004; Smith, 2005). Hay que evitar dejar fuera alguna faceta o dominio relevante (infrarrepresentación), así como ponderar en demasía una faceta o dominio de la variable (sobrerrepresentación) (Smith et al., 2003). Asimismo, no se deben incorporar facetas, o ítems, que no tengan relación con la variable objeto de medición (varianza irrelevante). Una definición operativa y precisa de la variable influye de forma determinante en la posterior obtención de los diferentes tipos de evidencias de validez, ayuda a especificar las conductas más representativas de la variable objeto de medición y facilita el proceso de construcción de los ítems (Carretero y Pérez, 2005; Elosua, 2003; Muñiz et al., 2005; Sireci, 1998a; Smith, 2005).

No solo es importante una definición operativa de la variable sino que también es preciso identificar y definir sus facetas o dominios (definición semántica) y la relación que se establece entre ellas, así como con otras variables de interés (definición sintáctica) (Lord y Novick, 1968). Lógicamente, las diferentes facetas que componen la variable medida se deberían encontrar relacionadas, dado que se supone que miden la misma variable o constructo. Al mismo tiempo hay que establecer la relación con otras variables de interés. La variable objeto de medición no se encuentra aislada en el mundo, sino que está en relación o interacción con otras variables. Es interesante comprender y analizar estas relaciones especificándolas de antemano con el propósito de llevar a cabo posteriores estudios dirigidos a la obtención de evidencias de validez (Carretero y Pérez, 2005; Muñiz, 2004; Smith, 2005).

3. ESPECIFICACIONES

Una vez delimitados el propósito de la evaluación y la definición operativa de la variable que interesa medir, se deben llevar a cabo determinadas especificaciones relacionadas con el instrumento de medida. En esta fase se deben describir de forma detallada y precisa aspectos concernientes a los requerimientos de aplicación del instrumento de medida, el tipo, número, longitud, contenido y distribución de los ítems, especificaciones e instrucciones en la entrega del material y aspectos relacionados con la seguridad del mismo.

Los requerimientos de aplicación del instrumento de medida se refieren a cuál va a ser el soporte de administración (papel o informático), a qué tipo de aplicación se va a realizar (individual y/o colectiva) y a cuándo y en qué lugar se va a administrar el instrumento de medida. Igualmente, se deben especificar los requerimientos cognitivos, de vocabulario y de accesibilidad de los participantes. Es importante llevar a cabo adaptaciones (acomodaciones) para aquellos participantes que no puedan desempeñar la tarea en igualdad de condiciones que el resto, por ejemplo, disponer de una versión en Braille para una persona con deficiencia visual. Las adaptaciones que se realicen deben estar convenientemente avaladas por evidencias empíricas para que no supongan ventajas ni desventajas respecto de la aplicación estándar (Dorans y Cook, 2016; Wells y Faulkner-Bond, 2016).

En relación con los ítems, se debe especificar su tipo, el número, la longitud, el contenido y el orden (disposición), así como el formato de respuesta o el tipo de alternativas que se van a utilizar. Con respecto a este tema, no existen normas universales, y todo dependerá de las circunstancias de aplicación, del propósito de la variable objeto de medición y de otras circunstancias.

4. CONSTRUCCIÓN DE LOS ÍTEMS

La construcción de los ítems constituye una de las etapas más cruciales dentro del proceso de elaboración del instrumento de medida. Los ítems son la materia prima, los ladrillos a partir de los cuales se conforma un instrumento de evaluación, por lo que una construcción deficiente de los mismos incidirá en las propiedades métricas finales del instrumento de medida y en la validez de las inferencias que se hagan a partir de las puntuaciones (Haladyna y Rodriguez, 2013; Lane, Raymond y Haladyna, 2016; Muñiz et al., 2005; Osterlind, 1998; Schmeiser y Welch, 2006).

Los principios básicos que deben regir la construcción de cualquier banco de ítems son: representatividad, relevancia, diversidad, claridad, sencillez y comprensibilidad (Muñiz et al., 2005). Todos los dominios de la variable de interés deben estar igualmente representados (evitando la infra o sobrerrepresentación), aproximadamente con el mismo número de ítems, a excepción de que se haya considerado un dominio más relevante dentro de la variable y que, por tanto, deba tener un mayor número de ítems, esto es, una mayor representación. Un muestreo erróneo del dominio objeto de evaluación sería una clara limitación en la obtención de evidencias de validez de contenido y tendrá repercusiones en las inferencias que con posterioridad se hagan a partir de las puntuaciones. Los ítems deben ser heterogéneos y variados para así recoger una mayor variabilidad y representatividad de la variable. Deben primar la claridad y la sencillez y se deben evitar tecnicismos, negaciones, dobles negaciones o enunciados excesivamente prolijos o ambiguos (Muñiz et al., 2005). Del mismo modo, los ítems deben ser comprensibles para la población a la cual

va dirigido el instrumento de medida, evitándose en todo momento un lenguaje ofensivo y/o discriminatorio. Ítems con una redacción defectuosa o excesivamente vagos van a incrementar el porcentaje de varianza explicada por factores espurios o irrelevantes, con la consiguiente merma en las evidencias de validez de la prueba.

Si los ítems provienen de otro instrumento ya existente en otro idioma y cultura, deberán seguirse las directrices internacionales para la traducción y adaptación de test (Hambleton, Merenda y Spielberger, 2005; Muñiz y Bartram, 2007; Muñiz, Elosua y Hambleton, 2013). En el caso de ítems originales, han de seguirse las directrices elaboradas para su desarrollo (Downing y Haladyna, 2006; Haladyna, 2004; Haladyna et al., 2002, 2013; Moreno et al., 2004, 2006, 2015; Muñiz et al., 2005).

Durante las fases iniciales de la construcción del banco de ítems se recomienda que el número de ítems inicial sea como mínimo el doble del que finalmente se considera que podría formar parte de la versión final del instrumento de medida. La razón es bien sencilla: muchos de ellos, por motivos diferentes (métricos, comprensibilidad, dificultad, etc.), se acabarán desechando, por lo que solo quedarán aquellos que ofrezcan mejores indicadores o garantías técnicas (sustantivas y métricas). Finalmente, para garantizar la obtención de evidencias de validez basadas en el contenido de los ítems, se ha de recurrir a la consulta de expertos y a la revisión exhaustiva de las fuentes bibliográficas, así como a otros instrumentos similares ya existentes (Sireci, 1998b; Sireci y Faulkner-Bond, 2014). En relación con la valoración de los ítems por parte de los expertos, y con la finalidad de una evaluación más precisa y objetiva del conjunto inicial de ítems, se puede pedir a los expertos que juzguen, a partir de un cuestionario, si los ítems están bien redactados para la población de interés, si son o no pertinentes para evaluar una faceta o dominio determinado y si cada ítem representa de forma adecuada la variable o dimensión de interés. A continuación se comentan algunos de los formatos más habituales de los ítems y se dan unas directrices para el desarrollo de ítems de elección múltiple, que son de los más utilizados, si no los más. Tratamientos exhaustivos pueden verse en Osterlind (1998), Haladyna y Rodríguez (2013) o Lane et al. (2016).

4.1. Tipos de ítems

Existe una gran variedad de ítems, que se pueden clasificar en distintas categorías en función de los criterios que se tengan en cuenta, tales como su contenido, el formato o la forma de respuesta exigida, bien sea seleccionar una respuesta entre las presentadas o desarrollarla (Downing, 2006; Haladyna y Rodríguez, 2013; Magno, 2009; Osterlind, 1998; Osterlind y Merz, 1994; Rauthmann, 2011; Sireci y Zenisky, 2006). Por ejemplo, Scalise y Gifford (2006) establecen siete tipos de ítems, que van desde la selección pura de la respuesta hasta la construcción completa, pasando por varias posibilidades intermedias. Por su parte, Sireci y Zenisky (2016) añaden todavía otros tipos de tareas. En suma, se han propuesto diversas clasificaciones, tratando de sistematizar y organizar la gran cantidad de tipos de ítems existentes, y si bien resultan útiles en la práctica, ninguna de ellas resulta totalmente satisfactoria desde un punto de vista teórico.

Esta proliferación de los tipos de ítems se ha acentuado en los últimos años debido a las grandes posibilidades que ofrecen las nuevas tecnologías de la información, que están influyendo de forma clara en su formulación (Sireci y Zenisky, 2016). Según Parshall et al. (2010), habría siete dimensiones o aspectos de los ítems en los que se están produciendo las mayores innovaciones debido a la irrupción de las nuevas tecnologías:

a) Su *estructura,* con la aparición de nuevos formatos facilitados por las nuevas tecnologías y las facilidades que ofrecen la pantallas de los ordenadores para su implementación.
b) *Complejidad,* al incluirse en los ítems nuevos elementos que han de tenerse en cuenta para responder.
c) *Fidelidad,* referida a la posibilidad que ofrecen las tecnologías de la información de dar un mayor realismo a los ítems.
d) *Interactividad,* dado que el ítem puede reaccionar y mutar en función de las respuestas de las personas, volviéndose interactivo.
e) *Multimedia,* cuando se incluyen en los ítems medios técnicos como audio, vídeo, gráficos, animación u otros.

f) *Tipo de respuesta,* habiendo una amplia gama de posibilidades del tipo de tareas que los ítems demandan.
g) *Sistemas de puntuación,* pudiendo registrarse, además de los clásicos aciertos y errores, otros muchos parámetros, tales como tiempos, intentos, estrategias, etc.

Un excelente trabajo sobre la influencia de los avances tecnológicos sobre los test puede verse en Drasgow (2016), y para el tema específico de la generación automática de ítems, Gierl y Haladyna (2013).

Aquí vamos a limitarnos a presentar algunos de los tipos de ítems más clásicos, remitiendo al lector interesado en saber más a la bibliografía citada más arriba.

Elección múltiple

Se trata del tipo de ítem más utilizado, y en su versión más genuina consta de un enunciado y varias alternativas, una de las cuales es la correcta. Dentro de los ítems selectivos, es decir, ítems en los que hay que seleccionar una opción, este formato es el más recomendable. Un ejemplo sería:

En los modelos de TRI la función de información es una medida de:

a) Fiabilidad.
b) Validez.
c) Eficiencia.

Recordará el lector que la función de información se refería a la fiabilidad; por tanto, la opción correcta es la *a*. Dado el uso masivo de este tipo de ítems en toda clase de ámbitos, en el subepígrafe 4.2 se presentan unas directrices para su construcción y evaluación.

Verdadero-falso

Se trata de un tipo de ítem también muy utilizado en el que la tarea de la persona evaluada consiste en decidir si una afirmación es verdadera o falsa. Por ejemplo:

— El valor de la sensibilidad de un test está entre 0 y 1.
— El coeficiente de validez es la correlación entre el test y el criterio.

En este caso ambas afirmaciones son correctas. Cuando se construye un test con este tipo de ítems, hay que equilibrar el número de ítems verdaderos y falsos. Este formato ha sido ampliamente estudiado. A continuación se sintetizan algunas de sus ventajas e inconvenientes, siguiendo a autores como Frisbie y Becker (1991) y Haladyna y Rodríguez (2013):

Ventajas

— Son fáciles de formular.
— Pueden utilizarse para evaluar todo tipo de contenidos.
— Se adaptan a diferentes demandas cognitivas.
— Para un mismo tiempo pueden hacerse muchos más que si fuesen de elección múltiple.
— Son fáciles de puntuar.
— Ocupan poco espacio, y resultan económicos.
— Son muy realistas, al plantear si una frase es verdadera o falsa.
— Reducen el tiempo de lectura.
— Los test compuestos por ellos dan buenos coeficientes de fiabilidad.

Desventajas

— Si no se hacen bien, tienden a centrarse en contenidos triviales, y esto hay que evitarlo. Ahora bien, más que un defecto del formato en sí, es del constructor, que no controla este aspecto.
— Tienden a ser muy memorísticos si no están bien construidos. De nuevo no se trata de un defecto del formato *per se,* sino del constructor.
— Los aciertos al azar son muy elevados, aunque este efecto puede mitigarse alargando el test.
— Dada su estructura dicotómica, no permite detectar fácilmente grados de verdad y falsedad.
— Los test compuestos por ítems verdadero-falso tienden a dar coeficientes de fiabilidad

algo más bajos que los test similares formados por ítems de elección múltiple.
— Según algunos autores, los ítems que son verdaderos no se comportan igual que los falsos.

En suma, los ítems verdadero-falso pueden utilizarse, son una opción, pero siguen siendo preferibles los de elección múltiple clásicos.

Elección múltiple compleja

Se trata de un formato en el cual se ofrecen distintas alternativas sobre las que se plantean varias opciones de respuesta. Por ejemplo:

¿Cuál de los siguientes indicadores constituye una evidencia de validez de un test?

1. Sensibilidad.
2. Especificidad.
3. Función de información.

a) 1 y 2.
b) 2 y 3.
c) 1 y 3.
d) 1, 2 y 3.

Recordará el lector, que la opción correcta es la *a*). Este formato se utiliza con cierta frecuencia, si bien Haladyna y Rodríguez (2013) no lo recomiendan por varias razones, entre ellas:

— Si la persona evaluada tiene un conocimiento parcial, por ejemplo, sabe que una determinada opción es correcta o incorrecta, puede ayudarle a encontrar la respuesta correcta eliminado distractores. Es decir, las habilidades y experiencia con los test pueden ayudar a una persona a encontrar la verdadera respuesta, interfiriendo con el nivel en la variable medida e introduciendo un sesgo indeseado.
— Con estos ítems se obtienen índices de discriminación más bajos, lo cual tiende a rebajar la fiabilidad del test.
— El formato resulta complejo de formular y exige bastante tiempo de lectura, reduciendo el número de ítems que se pueden incluir en un test. Esta reducción de ítems afecta al muestreo del constructo evaluado y por tanto a las evidencias de validez de contenido que se pueden obtener.

En suma, se trata de un formato poco recomendable.

Emparejamiento

En este conocido formato una serie de ideas, conceptos o hechos hay que emparejarlos con sus correspondientes siguiendo las directrices del enunciado, por ejemplo:

En el cuadro adjunto aparecen cuatro autores que debe relacionar con sus obras correspondientes.

1. Spearman	*a*) Coeficiente α.
2. Rasch	*b*) Modelo lineal clásico.
3. Cronbach	*c*) Modelo logístico de un parámetro.
4. Cohen	*d*) Coeficiente *kappa.*

Recordará el lector que los emparejamientos correctos serían: 1-*b*, 2-*c*, 3-*a*, 4-*d*.

Según Haladyna y Rodríguez (2013), las ventajas de este formato serían su fácil construcción, una presentación compacta, su popularidad y amplia aceptación y que permite evaluar procesos cognitivos superiores, tales como comprensión de conceptos, principios y procedimientos. Además, es muy eficiente en tiempo y espacio. Si bien es un formato muy querido en los reales de la enseñanza, en el caso de las variables psicológicas tiene el inconveniente de que, dada su naturaleza, no permite llevar a cabo un muestreo exhaustivo del constructo a evaluar, con los problemas que ello conlleva de cara a las evidencias de validez de contenido.

Likert

Este formato de los ítems, cuyo nombre proviene del trabajo original del autor (Likert, 1932), es tan omnipresente en el ámbito de la medición de las actitudes, opiniones, preferencias, creencias y otros campos afines no cognitivos como lo es el de elección múltiple para la medición de variables

cognitivas. En este formato la persona evaluada debe posicionarse en una escala de varias categorías ordinales entre las que tiene que elegir. Por ejemplo:

Me considero una persona ordenada:

1. Total desacuerdo.
2. En desacuerdo.
3. De acuerdo.
4. Totalmente de acuerdo.

La popularidad de este formato proviene de su facilidad de aplicación y de su adaptación a cualquier ámbito de evaluación, razón que explica en la práctica se haya impuesto a otros modelos mejor fundados científicamente, pero de aplicación menos sencilla, como el de las comparaciones binarias de Thurstone (1927a, 1927b, 1928b), entre otros.

Hay una literatura abundante sobre su construcción y uso (Dillman et al., 2009; Haladyna y Rodríguez, 2013; Krosnick y Presser, 2010). Comentaremos aquí los aspectos más relevantes, relativos a su formulación y al número de categorías.

— La frase sobre la que hay que pronunciarse debe estar claramente formulada, con un lenguaje claro y directo, sin ambigüedad, de modo que sepamos con precisión sobre qué se está posicionando exactamente la persona evaluada.
— No hay que utilizar ítems que se refieren a dos cuestiones al mismo tiempo, para evitar confusiones. Por ejemplo, si utilizamos el ítem *me gusta el pan y el vino,* nunca sabremos con precisión sobre qué se está pronunciando la persona evaluada, de modo que mejor desglosar el ítem en dos, uno para el pan y otro para el vino.
— Utilizar un lenguaje adaptado a la población a la que va dirigida la prueba, evitando tecnicismos o palabras rebuscadas.
— Utilizar frases cortas, evitando la verbosidad innecesaria.
— Evitar que se produzcan efectos suelo y techo, es decir, que la mayoría o todas las respuestas de las personas evaluadas se concentren en la categoría más baja (suelo), o en la más alta (techo). Si en el análisis de los ítems se detectan este tipo de ítems, hay que eliminarlos de la prueba o reformularlos para futuras evaluaciones.
— Formular los ítems de forma positiva, evitar los negativos. Este punto es polémico, pues hay autores que recomiendan balancear los positivos y negativos y luego recodificar los negativos para obtener la puntuación total. Esta recomendación de balancear los ítems positivos y negativos viene motivada por la creencia de que así se podría evitar la aquiescencia a la hora de responder, pero tiene más inconvenientes que ventajas, por lo que se desaconseja su uso. Por un lado, no está nada claro que de hecho evite o mitigue la aquiescencia, y, por otro, la recodificación de los ítems es una práctica muy problemática, por las asunciones en las que se basa.
— El número de alternativas que debe tener un ítem tipo Likert es una cuestión ampliamente investigada, y que ofrece muchos matices. A continuación se resumen los resultados más destacados. No deben utilizarse ítems con tres alternativas, pues la alternativa central tiende a atraer una gran cantidad de respuestas, lo que rebaja la capacidad discriminativa del ítem. La mayoría de las investigaciones indican que los ítems tipo Likert funcionan bien cuando se utilizan entre cuatro y siete categorías, siendo muy frecuente el uso de cinco. Podría argumentarse que en el caso de cinco la alternativa central podría monopolizar las respuestas, como ocurría con tres, pero su efecto es menor al existir otras cuatro opciones. Con cuatro o cinco categorías los ítems funcionan bien y tiene la ventaja de que resulta sencillo asignar un nombre o etiqueta a cada categoría, lo cual se vuelve más difícil si se aumenta el número de categorías. En algunas poblaciones familiarizadas con determinadas escalas podrían utilizarse estas, como es el caso de las poblaciones de estudiantes, que conocen perfectamente la escala 0-10 utilizada para sus calificaciones. La ventaja de estas escalas más amplias es que tienden a aumentar la variabilidad de las puntuaciones de las

personas, lo cual siempre es deseable de cara a los análisis estadísticos y psicométricos posteriores. Algunos autores no aconsejan este uso (Couper et al., 2006; Krosnick, 1999), sobre todo con poblaciones poco familiarizadas con ellas, por la posibilidad de que generen patrones de respuesta diferenciales en función de la interpretación de la escala numérica, no anclada a una denominación concreta de las categorías. Un análisis detallado de la influencia del número de categorías de los ítems tipo Likert sobre las propiedades psicométricas de la escala puede verse en Muñiz et al. (2005) y Lozano, García-Cueto y Muñiz (2008). Para un análisis de los errores en el uso de este formato véase Carifio y Perla (2007).

Ensayo

En todos los formatos de los ítems vistos más arriba la persona tenía que elegir o seleccionar una respuesta entre las que se le ofrecían; por eso a veces se alude a ellos con la expresión general de «formatos selectivos», o «de elección». Sin embargo, en el formato que se comenta ahora, denominado «de ensayo», «desarrollo» o «construcción», la persona evaluada tiene que construir la respuesta. Aunque existen posibles variaciones, aquí se comenta el formato genuino en el que se pide a la persona evaluada que desarrolle un determinado tema, por ejemplo el *concepto y tipos de fiabilidad.* Este formato es más utilizado en el ámbito educativo que en el estrictamente psicológico, pero es aplicable a cualquier contexto. La gran ventaja de los formatos de construcción frente a los selectivos es que permite una mayor libertad de expresión de la persona evaluada, pudiendo apreciarse su capacidad de expresión, su creatividad, su estilo y organización, amén de su dominio del tema propuesto. Por estas razones es un formato muy apreciado entre los educadores, que con cierta frecuencia lo prefieren a los formatos de elección, los cuales, más que generar la propia respuesta, exigen a la persona evaluada reconocer entre las alternativas propuestas la correcta. Pero todo tiene un precio, y el de los formatos de desarrollo es la posible subjetividad a la hora de la corrección y puntuación, lo cual hay que evitarlo a toda costa, por razones obvias. Por ejemplo, cuando la respuesta se escribe a mano, como es muy habitual, el tipo de letra puede condicionar notablemente al corrector, perjudicando claramente a las personas con peor letra. Este aspecto podría evitarse utilizando un procesador de texto, cuando ello sea posible. El efecto de halo es otro posible sesgo, dado que la opinión que el evaluador tiene sobre la persona evaluada puede condicionar por generalización la corrección. La solución es eliminar los nombres de las personas de la prueba antes de llevar a cabo una corrección para que esta sea ciega respecto de la persona evaluada. El efecto anclaje puede ser otro condicionante de la corrección; por ejemplo, si el evaluador acaba de corregir una prueba con un gran nivel, sin pretenderlo va a tender a evaluar con más rigor a la siguiente persona que corrija al quedar anclado al elevado nivel de la anterior. Para evitar estos y otros sesgos es muy importante instruir y entrenar de forma adecuada a los correctores, así como enseñarles a establecer unos criterios claros que les permitan una corrección más analítica y objetiva. Estos criterios de corrección, denominados «rúbricas», no solo permiten una mayor objetividad a la hora de corregir, sino que han de hacerse públicos para orientar a las personas evaluadas. Por ejemplo, para la evaluación del trabajo fin de grado (TFG) en la Facultad de Psicología de la Universidad de Oviedo, se establecen los siguientes criterios con sus correspondientes pesos, y además cada uno de esos criterios o dimensiones se desglosa con mayor detalle, tratando de evitar ambigüedades y evaluaciones no equitativas:

Aspectos formales (0-10 puntos).

— Estructura del trabajo: 0-4
— Ajuste a normas APA: 0-3
— Redacción: 0-3

Contenidos (0-45 puntos).

— Fundamentación: 0-10
— Originalidad y objetivos: 0-10
— Metodología: 0-10
— Resultados y conclusiones: 0-10
— Bibliografía: 0-05

Defensa pública (0-30 puntos).

— Estructura y calidad técnica: 0-10
— Claridad expositiva: 0-10
— Interacción con el tribunal: 0-10

Valoración del tutor (0-15 puntos).

— Asistencia a las tutorías: 0-3
— Seguimiento de las orientaciones: 0-3
— Cumplimiento plan de trabajo: 0-3
— Interés e implicación en el TFG: 0-3
— Entrega dentro del plazo establecido: 0-3

Estas directrices o rúbricas no son la panacea, pero ayudan a objetivar la evaluación y a evitar sesgos y subjetivismos. El problema de la construcción y valoración de los ítems de ensayo está muy ampliamente tratado en la bibliografía psicométrica; pueden consultarse, por ejemplo, las directrices del Educational Testing Service (Baldwin et al., 2005; Livingston, 2009), Hogan y Murphy (2007) y un buen resumen en Haladyna y Rodríguez (2013). Tal vez la solución radical a la objetividad de la evaluación de los ensayos venga a través de una vigorosa línea actual de investigación psicométrica sobre la corrección automática mediante programas de ordenador. Puede sorprender al lector la posibilidad de que un ensayo pueda ser corregido por un programa informático, pero los avances en este campo son notables, existiendo ya programas con altas prestaciones (Livingston, 2009; Shermis y Burstein, 2013; Williamson et al., 2006, 2010).

4.2. Directrices para la construcción de ítems de elección múltiple

Los ítems de elección múltiple son los más utilizados en numerosos ámbitos de la evaluación psicológica, educativa y en general en las ciencias sociales y de la salud. Su construcción inadecuada puede causar serios perjuicios a las personas evaluadas y atentar contra la equidad de los test. Se presentan a continuación las directrices para su construcción, desarrolladas por Haladyna et al. (2002, 2013) y Rodríguez (2016). Se refieren al contenido, formato, estilo, redacción del enunciado y redacción de las alternativas de los ítems.

Contenido

1. *Cada ítem debe centrarse en un contenido específico y en una determinada demanda cognitiva.*

   Cada ítem irá dirigido a evaluar un contenido y solo uno, sin mezcla ni confusión, elaborando todos los ítems que fuesen necesarios para muestrear adecuadamente los contenidos. Asimismo, el ítem deberá centrarse en un solo proceso cognitivo, como puede ser comprender un concepto, aplicar un conocimiento, hacer inferencias o lo que fuere.
2. *Utilizar material novedoso para evaluar el aprendizaje de alto nivel.*

   Cuando se utilice parte de un libro, texto o documento en la evaluación, conviene parafrasearlo, no utilizarlo literalmente, para así evitar evaluar el mero recuerdo. Lo mismo debe hacerse con cualquier tipo de materiales utilizados en la enseñanza.
3. *Mantener el contenido de cada ítem independiente del contenido de otros ítems del test, no utilizar ítems encadenados.*

   Un ítem no tiene que hacer alusión a ningún otro ni estar encadenado a los anteriores. Cuando se utiliza un texto o material sobre el cual se hacen varios ítems, hay que ser especialmente cuidadoso con que unos ítems hagan alusión a otros. Si esto ocurriera, según Haladyna y Rodríguez (2013), tiende a sobreestimarse la fiabilidad. Además, tienden a beneficiar a las personas con experiencia en hacer los test, aunque no controlen la materia evaluada. Recuérdese que los modelos psicométricos de TRI asumen la independencia local, es decir, que la respuesta a un ítem no influye en la respuesta a otro, con lo que se estaría violando este principio.
4. *El ítem debe evaluar un contenido importante, evitando contenidos muy específicos o muy generales.*

   El número de ítems de un test es limitado, así que hay que centrarse en los temas más relevantes a evaluar, evitando cuestiones marginales o muy generales.

5. *Evitar ítems cuyas opciones sean opinables.*
   No se pueden utilizar ítems cuyas opciones no sean claras y planteen dudas entre los expertos. Las opciones tienen que ser claras, unívocas e indiscutibles, cartesianas, pues de otro modo se estaría introduciendo una gran arbitrariedad en la evaluación, lo que podría llevar a anular a posteriori los ítems con alternativas discutibles. Esto ocurre con frecuencia en los exámenes para psicólogos internos residentes (PIR) y médicos (MIR). Desde luego, anular a posteriori un ítem dudoso es un mal menor, pero es un mal, pues algunos candidatos lo habrán acertado, y se les anula, y otros habrán invertido una gran cantidad de tiempo en resolverlo en detrimento de otros ítems. Si la prueba se construye con rigor, esto no debería ocurrir, pues demuestra una falta de profesionalidad con consecuencias no deseables para las personas evaluadas.
6. *Evitar ítems con trampas.*
   No se trata de construir ítems para ir a «cazar» a las personas evaluadas mediante ítems truculentos, sino de comprobar su nivel de conocimientos de manera clara, directa y profesional. Además, si se utilizan este tipo de ítems, se genera una actitud de desconfianza de las personas evaluadas en el proceso de evaluación, lo cual no es deseable.

Formato

7. *Formular las alternativas del ítem de forma vertical, no horizontal.*
   La única ventaja de formular las alternativas de forma horizontal es que se ahorra espacio y la edición de la prueba resulta más barata, pero tiene muchos inconvenientes, pues resulta más difícil de leer y en consecuencia aumenta los errores y confusiones innecesariamente; por tanto, es mejor utilizar siempre el formato vertical, por ejemplo:

   El estudio de la precisión de las puntuaciones de un test se denomina:

   a) Validez.
   b) Fiabilidad.
   c) Variabilidad.

   Como habrá adivinado el lector, la respuesta correcta es la *b*).

Estilo

8. *Corregir y probar los ítems antes de llegar a la versión definitiva.*
   Hay que asegurarse de que los ítems están formulados de forma correcta, tanto en lo relativo a los aspectos gramaticales y ortográficos como en el contenido. Aparte de lo embarazoso y poco profesional que puede resultar para quienes aplican los test tener que hacer rectificaciones en la sesión de evaluación, una formulación incorrecta del ítem afecta a la validez de las inferencias hechas a partir de las puntuaciones del test. En suma, hay que revisar con sumo cuidado los ítems antes de incluirlos en un test.
9. *Ajustar la complejidad del lenguaje a las personas evaluadas.*
   La idea central es que las personas evaluadas deben de comprender el ítem, otra cosa es que luego conozcan o no la repuesta. Si no lo comprenden, no podremos saber si los fallos cometidos se deben a que desconocen las respuestas o que no han comprendido el enunciado. Las personas evaluadas tienen que comprender bien los ítems, la comprensión lectora no debe de interferir con la variable medida
10. *Minimizar la cantidad de lectura en cada ítem.*
    Ya nos lo dejó dicho nuestro Baltasar Gracián: lo bueno, si breve, dos veces bueno. Hay que evitar enunciados farragosos e innecesariamente largos que interfieren con la variable que queremos evaluar. Esto ocurre con cierta frecuencia en la evaluación educativa cuando se utilizan enunciados alambicados, en los que es difícil saber cuál es exactamente lo que se pregunta. Por ejemplo, si en un problema de matemáticas utilizamos una formulación muy

compleja, al final no sabemos si los estudiantes desconocen el problema matemático planteado o sencillamente no tienen claro qué se pregunta. Además, al reducir la longitud de los enunciados, podemos incluir más ítems en el mismo tiempo destinado a la evaluación, con el consiguiente beneficio que ello supone para la fiabilidad y validez del test, tal como se vio en los apartados correspondientes.

Redacción del enunciado

11. *Incluir la idea central en el enunciado y no en las alternativas.*

    La parte importante del texto debe ir en el enunciado, dejando las alternativas con un texto breve. Un error bastante común es hacer un enunciado breve y luego meter demasiado texto en las alternativas, lo cual aumenta innecesariamente la longitud del ítem y hace la tarea de la persona evaluada más difícil, pues no conoce el objetivo final del ítem hasta que no lee todas las alternativas que contienen la mayor parte de la información.

12. *Expresar el enunciado en términos positivos.*

    Hay que evitar formular los ítems de forma negativa. La razón de esta recomendación es que a las personas les resulta mucho más difícil comprender textos formulados de forma negativa que positiva. Por tanto, al formularlo negativamente, la comprensión del ítem estaría interfiriendo con la variable medida. Por ejemplo, si pretendo evaluar la competencia matemática de las personas, no es deseable que desempeñe un papel importante la comprensión verbal, ya que distorsionaría el resultado.

Redacción de las alernativas

13. *Utilizar opciones que sean plausibles y discriminativas.*

    Las alternativas del ítem tienen que tener sentido, ser plausibles y discriminar entre las personas más competentes en la variable medida y las menos competentes. Hay que evitar alternativas obvias o absurdas que rebajan la discriminación del ítem. En el apartado dedicado al análisis de los ítems ya se expuso cómo se pueden evaluar empíricamente estos aspectos una vez que se ha aplicado el test. También vimos allí los argumentos psicométricos que indican que el número aconsejable de alternativas es tres, frente a cuatro o cinco.

14. *Asegurarse de que solo una de las alternativas es la respuesta correcta.*

    Puede parecer una obviedad, pero conviene asegurarse de que no hay dudas sobre cuál es la respuesta correcta. Esto se consigue con una buena revisión por parte de varios expertos; si hubiese alguna duda al respecto, hay que eliminar o corregir el ítem. Si se detecta el problema tras llevar a cabo la evaluación, hay que suprimir el ítem como mal menor, pero, como se indicó más arriba, no es deseable, pues plantea serios problemas deontológicos al perjudicar a las personas evaluadas.

15. *Variar la colocación de la respuesta correcta.*

    Se refiere a que hay que repartir aleatoriamente la alternativa en la que se coloca la respuesta correcta. Si, por ejemplo, un test consta de treinta ítems de tres alternativas cada uno, la respuesta correcta debería aparecer diez veces en la primera alternativa, diez en la segunda y diez en la tercera. Hay que asegurarse de que esto se lleva a cabo correctamente, pues si se deja al mero azar del constructor del test podría ocurrir perfectamente que por determinada querencia no consciente acabase, por ejemplo, apareciendo la respuesta correcta en primer lugar en el 80 % de los casos, lo cual podría ayudar a las personas evaluadas a obtener aciertos de forma espuria.

16. *Colocar las alternativas en un orden lógico o numérico.*

    Cuando en las alternativas del ítem aparezcan cantidades numéricas, o ciertos argumentos lógicos, hay que presentarlos ordenados. Los ítems tratan de detectar si la persona evaluada domina determinada

materia, y el hecho de que los materiales se presenten desordenados podría contribuir a la confusión, interfiriendo con el verdadero nivel de la persona en la variable medida. El ordenamiento en las alternativas puede ser ascendente o descendente; por ejemplo en el siguiente ejemplo se plantean en orden ascendente:

¿En qué año se publicó el libro de Rasch sobre el modelo logístico de un parámetro?

a) 1954.
b) 1960.
c) 1975.

Recordará el lector que la correcta es la opción *b*).

17. *Construir las alternativas independientes entre sí, no deben solaparse.*

Cada alternativa de un ítem tiene vida propia, debe ser independiente del resto de alternativas, sin solapamientos con las demás. Si resultan interdependientes, pueden dar claves a las personas evaluadas sobre la opción correcta, aparte de generar confusión.

18. *No deben utilizarse las alternativas: ninguna de las anteriores, todas las anteriores, no lo sé.*

Desgraciadamente es muy frecuente encontrar evaluaciones, sobre todo educativas, en las que se incluyen este tipo de opciones, que son totalmente desaconsejables. La razón fundamental de su inadecuación es que van a permitir a las personas evaluadas hallar pistas sobre la alternativa correcta, independientemente de su nivel en la variable medida. Es decir, a igual nivel de competencia en la variable medida, estas alternativas van a favorecer a las personas más hábiles o avezadas en la práctica de los test, desvirtuando así la medida genuina de la variable objeto de la medición.

Es verdad que algunos autores mantienen abierta la posibilidad de utilizar la opción *ninguna de las anteriores,* pero nos unimos a quienes mantienen su inoportunidad, pues si la pregunta que se plantea en un ítem tiene respuesta correcta, ¿por qué no ha de figurar en el ítem? Parece la forma más clara, directa y honesta intelectualmente de proceder. Como señalan Haladyna y Rodríguez (2013), no existe ninguna ventaja en omitir la respuesta correcta en la formulación del ítem.

En el caso de la opción *todas las anteriores,* el mayor problema es que puede dar pistas a las personas evaluadas sobre la opción correcta, favoreciendo a aquellas personas más experimentadas en la práctica de los test. Por ejemplo, si en un ítem de cuatro alternativas, siendo la última *todas las anteriores,* una persona sabe que dos de las tres restantes son correctas, aun desconociendo la tercera, infiere acertadamente que la correcta será *todas las anteriores,* pues el ítem no puede tener dos correctas; luego solo cabe que lo sean todas. En suma, la habilidad de la persona para hacer test está interactuando con su nivel en la variable medida, lo cual no es deseable.

En el caso de *no lo sé,* tampoco mejora en nada la evaluación, y es mejor evitarla. No todas las personas evaluadas la interpretan de la misma manera, por lo que puede introducir sesgos indeseables en la evaluación, de modo que es mejor evitarla.

19. *Formular las alternativas de forma positiva.*

Como ya se indicó en la directriz 12 para el enunciado del ítem, en el caso de las alternativas también deben formularse de forma afirmativa por las dificultades que conlleva para las personas evaluadas el procesamiento de frases negativas. Como ocurría en el enunciado, el uso de términos como NO y EXCEPTO debe evitarse en la redacción de las alternativas.

20. *Evitar dar pistas sobre la respuesta correcta.*

a) *Longitud de las alternativas.*

Las alternativas deben tener aproximadamente la misma longitud. Un error muy frecuente es redactar de forma larga y detallada la opción correcta y luego despachar de forma breve

las incorrectas, dando la pista a las personas evaluadas de que la correcta es la opción larga. Hay que evitar este error a toda costa.

b) *Determinantes específicos.*

Hay que evitar dar pistas mediante la utilización de determinantes tales como *siempre, nunca, totalmente, absolutamente* o *completamente.* Estos determinantes son tan extremos que raramente van a figurar en la respuesta correcta, por lo que están indicando a las personas evaluadas que la opción en la que aparecen no es la correcta. Lo mejor es evitarlos.

c) *Asociaciones de términos.*

A veces en la formulación de los ítems se pueden introducir términos en el enunciado asociados con alguna de las alternativas, dando pistas sobre la respuesta correcta. Hay que evitar estas asociaciones, que pueden ser de sonidos similares, términos idénticos en el enunciado y las alternativas u otras posibilidades. Por ejemplo, el siguiente ítem estaría cayendo en este error, al incluir el término «fiabilidad» en el enunciado y en la tercera alternativa, que es la correcta, dando pistas obvias sobre ella:

La fiabilidad de las puntuaciones de un test se evalúa mediante el coeficiente de:

— Determinación.
— Validez.
— Fiabilidad.

d) *Pares o tríos de alternativas que den pistas sobre la opción correcta.*

Hay que evitar incluir entre las alternativas pares o tríos de ellas que por su estrecha relación den pistas claras a las personas evaluadas sobre cuál es la respuesta correcta.

e) *Alternativas claramente absurdas o ridículas.*

Con demasiada frecuencia se encuentran ítems que incluyen entre sus alternativas alguna que es absurda, ridícula, de dudoso gusto y por tanto obvia para todas las personas evaluadas. Estas alternativas no contribuyen a la calidad del ítem, por lo que no deben utilizarse en ningún caso, son alternativas perdidas. El siguiente ítem sería un ejemplo claro de lo dicho:

El modelo lineal clásico fue formulado originalmente por:

— Spearman.
— Obama.
— Cervantes.

Una persona puede desconocer que fue Spearman, quien en 1904 formulase el modelo lineal clásico, pero seguro que no fallará el ítem, dada su formulación lamentable.

f) *Homogeneidad de las alternativas.*

Si las alternativas son muy heterogéneas en contenido o en su estructura gramatical pueden dar pistas sobre la respuesta correcta, conviene que tengan una cierta homogeneidad. Un trabajo detallado sobre este aspecto puede consultarse en Ascalon et al. (2007).

21. *Hacer plausibles todas las alternativas incorrectas.*

Las alternativas incorrectas, denominadas con frecuencia «distractores», han de ser todas plausibles a priori. Es decir, todas las alternativas incorrectas deben tener su lógica y permitir la discriminación entre las personas más competentes y menos competentes en la variable medida. En contextos educativos pueden utilizarse los errores más comunes de los estudiantes en la formulación de los distractores.

22. *No utilizar el humor.*

No debe utilizarse el humor ni las gracias en la formulación de los ítems. Por supuesto, el uso del humor puede ser un

recurso interesante en numerosas situaciones de enseñanza y aprendizaje, contribuyendo a rebajar las tensiones y a crear un buen clima, pero su uso en la formulación de los ítems no contribuye a mejorar la evaluación, todo lo contrario. Como señalan Haladyna y Rodríguez (2013), supone el desperdicio del distractor en el que se incluye, haciendo, por tanto, el ítem más fácil. También puede contribuir a que las personas evaluadas no se tomen en serio la evaluación, y en algunos casos puede generar reacciones negativas de las personas evaluadas, sometidas como están a la presión de la evaluación.

5. EDICIÓN

En esta fase se compone y se imprime la primera versión del test, además de construir la base de datos con las claves de corrección. Este paso ha sido con frecuencia injustamente infraestimado y, sin embargo es esencial, pues el continente bien podría echar a perder el contenido. Buenos ítems pobremente editados dan como resultado un mal test, igual que los malos barriles pueden echar a perder los buenos vinos. Podemos haber construido un buen banco de ítems que de nada servirá si luego estos se presentan de forma desorganizada, con errores tipográficos, o en un cuadernillo defectuoso. Uno de los errores más frecuentes entre los constructores de test aficionados es utilizar fotocopias malamente grapadas, con la excusa de que solo se trata de una versión experimental de la prueba, olvidándose de que para las personas que las responden no existen pruebas experimentales, todas son definitivas. El aspecto físico de la prueba forma parte de la validez aparente. Es importante que el instrumento dé la impresión de medir de manera objetiva, rigurosa, fiable y válida la variable de interés, porque, entre otros aspectos, influye en un punto esencial presente en todo el proceso de evaluación: la motivación de las personas evaluadas. Por otra parte, en esta fase también se debe construir, si fuera el caso, la base de datos donde posteriormente se van a tabular las puntuaciones y a realizar los análisis estadísticos pertinentes, así como las normas de corrección y puntuación, por ejemplo si existen ítems que se deben recodificar, si se va a crear una puntuación total o varias puntaciones, etcétera.

6. ESTUDIOS PILOTO

La finalidad de cualquier estudio piloto es examinar el funcionamiento general del instrumento de medida en una muestra de participantes con características semejantes a la población objeto de interés. Esta fase es de suma importancia, ya que permite detectar, evitar y corregir posibles errores, así como llevar a cabo una primera comprobación del funcionamiento del instrumento de evaluación en el contexto aplicado. El estudio piloto podría verse como una representación en miniatura de lo que posteriormente va a ser el estudio de campo.

Existen dos tipos fundamentales de estudio piloto: cualitativo y cuantitativo (Wilson, 2005). El estudio piloto cualitativo permite, a partir de grupos de discusión, debatir diferentes aspectos relacionados con el instrumento de medida, por ejemplo la detección de errores semánticos o gramaticales, el grado de comprensibilidad de los ítems, las posibles incongruencias semánticas, etc. Los participantes en este pilotaje pueden ser (o no) similares a la población objeto de medición. Por su parte, el estudio piloto cuantitativo permite examinar las propiedades métricas de la versión preliminar del instrumento de medida, y ha de llevarse a cabo con personas similares a las que va dirigida la prueba. En ambos casos se deben anotar de forma detallada todas las posibles incidencias acaecidas durante la aplicación, por ejemplo preguntas o sugerencias de los participantes, grado de comprensión de los ítems, así como posibles errores o problemas detectados en el instrumento.

A continuación, una vez tabulados los datos, se procede a los análisis de la calidad psicométrica de los ítems. En función de criterios sustantivos y estadísticos, algunos ítems se mantienen mientras que otros son descartados o modificados. Es importante que el constructor del instrumento de evaluación deje constancia de qué ítems fueron eliminados o modificados y por qué, además de explicitar con claridad el criterio (cualitativo o cuantitativo) por el cual se eliminaron. En este paso, si se considera

conveniente, se pueden incorporar nuevos ítems. Todas las actividades deben ir destinadas a seleccionar los ítems con mayores garantías métricas que maximicen las propiedades finales del instrumento de evaluación. Finalmente, se debe construir una nueva versión del instrumento de medida que es revisada de nuevo por el grupo de expertos y que será la que en última instancia se administre en el estudio final de campo.

7. SELECCIÓN DE OTROS INSTRUMENTOS DE MEDIDA

La selección adecuada de otros instrumentos de evaluación permite recoger evidencias a favor de la validez de las puntuaciones de los participantes (Elosua, 2003). Es interesante que no se pierda el norte: la finalidad última de todo proceso de construcción de instrumentos de medida es siempre obtener evidencias de validez. La selección adecuada de otras variables de interés permite aglutinar diferentes tipos de evidencias que conduzcan a una mejor interpretación de las puntuaciones en el instrumento de medida dentro de un contexto y uso particular. En este sentido, se pueden establecer relaciones con un criterio externo, con otros instrumentos de medida que pretendan medir la misma variable u otras diferentes (lo que anteriormente se había denominado «definición sintáctica»). Las asociaciones entre las variables son la base para la obtención de evidencias de validez de relación con variables externas, que permite la construcción de una red nomológica.

La decisión de qué instrumentos se deben utilizar complementariamente con el nuestro viene afectada tanto por cuestiones sustantivas como pragmáticas, referidas a exigencias de tiempo y lugar y, cómo no, materiales (por ejemplo, posibilidad de acceso al test, cuestiones económicas, etc.). Evidentemente las exigencias materiales y temporales, así como las razones éticas, no permiten administrar todos los instrumentos que quisiéramos, si bien aquí no se trata de pasar cuantos más mejor, sino de seleccionar aquellos de mayor calidad científica, a partir de los cuales se pueda profundizar en el significado de nuestras puntuaciones. Algunas recomendaciones prácticas en la selección de otros instrumentos de medida son:

a) Que se encuentren validados para la población objeto de interés y se conozcan sus propiedades psicométricas.
b) Que sean sencillos y de rápida administración.
c) Que tengan «coherencia» sustantiva de cara a establecer relaciones entre las variables, dentro de una red nomológica.

8. APLICACIÓN DEL TEST

En esta fase de estudio de campo se incluyen la selección de la muestra (tipo, tamaño y procedimiento), la aplicación propiamente dicha del instrumento de medida a los participantes y el control de calidad y seguridad de la base de datos.

La representatividad y generalizabilidad de los resultados dependen en gran medida de que la muestra elegida sea realmente representativa de la población objetivo de estudio. Elegir una muestra pertinente en cuanto a representatividad y tamaño es esencial: si se falla en esto, todo lo demás va a quedar invalidado. El muestreo probabilístico siempre es preferible al no probabilístico; para la estimación del tamaño muestral requerido para un determinado error de medida ha de acudirse a los textos especializados, o consultar a los expertos en la tecnología de muestreo. Aunque no hay recetas universales y aún no se dispone de una base sólida para tal afirmación, se suele recomendar que por cada ítem administrado tengamos al menos cinco o 10 personas, o unas 200 observaciones como mínimo (Ferrando y Anguiano, 2010), si bien determinadas técnicas estadísticas pueden reclamar incluso más de cara a una buena estimación de los parámetros, por ejemplo los modelos de teoría de respuesta a los ítems.

Las actividades relacionadas con la aplicación y el uso del instrumento de medida, son cruciales durante el proceso de validación (Muñiz y Bartram, 2007; Muñiz et al., 2005). Cuando aplicamos cualquier instrumento de medida, hay que cuidarse de que las condiciones físicas de la aplicación sean las adecuadas (luz, temperatura, ruido, comodidad de los asientos, etc.). Igualmente, las personas encargadas de la administración del instrumento de medida deben establecer una buena relación *(rapport)* con los participantes, estar familiarizados con la

administración de este tipo de herramientas, dar las instrucciones a los participantes correctamente, ejemplificar con claridad cómo se resuelven las preguntas, supervisar la administración y minimizar al máximo las posibles fuentes de error. Por todo ello es recomendable elaborar unas pautas o directrices que permitan estandarizar la administración del instrumento de medida y garanticen la equidad.

El control de calidad de la base de datos es otro tema a veces poco valorado en el proceso de construcción de instrumentos de medida. Por control de calidad nos referimos a una actividad que tiene como intención comprobar que los datos introducidos en la base de datos se correspondan exactamente con las puntuaciones de los participantes en la prueba. Frecuentemente, cuando introducimos las puntuaciones de los participantes en una base de datos se pueden cometer multitud de errores, razón por la cual es altamente recomendable comprobar de forma rigurosa que los datos se han introducido correctamente. Una estrategia sencilla que se puede utilizar a posteriori es la de extraer al azar un cierto porcentaje de los participantes y comprobar la correspondencia entre las puntuaciones en la prueba y la base de datos. No obstante, los mejores errores son los que no se cometen, así que hay que poner todos los medios para minimizar los errores a la hora de construir la base de datos.

9. PROPIEDADES PSICOMÉTRICAS

Una vez aplicado el test a la muestra de interés, se procede al estudio de las propiedades psicométricas de sus puntuaciones: análisis de los ítems, estimación de la fiabilidad de las puntuaciones, obtención de evidencias de validez (por ejemplo, estudio de la dimensionalidad, análisis del funcionamiento diferencial de los ítems, relación con variables externas) y construcción de baremos. Como se ha visto con detalle en el epígrafe 2, la fiabilidad se refiere a la precisión de las puntuaciones, esto es, a la calidad de los datos, mientras que la validez se refiere a la calidad de las inferencias hechas a partir de las puntuaciones (Prieto y Delgado, 2010). En sentido estricto no es fiable el test, sino las puntuaciones obtenidas en él. Análogamente, un test no es válido, sino que lo son las inferencias hechas a partir de las puntuaciones. Nótese que todo proceso de medición en la ciencia que sea, la psicología incluida, conlleva necesariamente un error, el cual tiene que quedar claramente reflejado cuando el profesional lleva a cabo una evaluación.

En esta fase debe primar por encima de todo el rigor metodológico. Todos los pasos y decisiones que se tomen se deben describir con claridad y deben estar correctamente razonadas. En un primer lugar se deben analizar los ítems tanto a nivel cualitativo como cuantitativo. Para seleccionar los mejores ítems desde el punto de vista psicométrico se pueden tener en cuenta el índice de dificultad (cuando proceda), el índice de discriminación, las cargas factoriales y/o el funcionamiento diferencial de los ítems (Muñiz et al., 2005). El funcionamiento diferencial de los ítems trata de garantizar la equidad en el proceso de medición. La ausencia de funcionamiento diferencial en un ítem supone que la probabilidad de respuesta correcta depende únicamente del nivel del participante en la variable objeto de medición, y no está condicionada por la pertenencia a un grupo determinado o característica, por ejemplo, género, cultura u otro aspecto cualquiera (Gómez, Hidalgo y Guilera, 2010). No se debe perder de vista que la finalidad del análisis psicométrico de los ítems es maximizar o potenciar las propiedades métricas del instrumento de medida; no obstante, no existen reglas universales, y las consideraciones estadísticas no garantizan unos resultados con significación conceptual, por lo que hay que tener presente también los aspectos sustantivos (Muñiz et al., 2005).

Una vez seleccionados los ítems, se procede al estudio de la dimensionalidad del instrumento para obtener evidencias de validez de su estructura interna. En el caso de encontrar una solución esencialmente unidimensional, nos podríamos plantear la construcción de una puntuación total, y en el caso de una estructura multidimensional deberíamos pensar en un conjunto de escalas o perfil de puntuaciones. El análisis factorial exploratorio y confirmatorio y el análisis de componentes principales son las técnicas multivariantes más utilizadas para examinar la estructura interna que subyace a las puntuaciones de un instrumento de evaluación (Ferrando y Anguiano, 2010), si bien no son las únicas (Cuesta, 1996). Una vez determinada la dimensionalidad de las puntuaciones del instrumento de medida, se lleva a cabo una estimación de la fiabilidad,

para lo cual se pueden seguir diversas estrategias, tanto desde el punto de vista de la teoría clásica de los test como de la teoría de respuesta a los ítems (Muñiz, 1997a, 2003). Posteriormente, y de cara a obtener evidencias de validez, se debe observar la relación del instrumento de medida con otros instrumentos de evaluación, y finalmente se lleva a cabo una baremación del instrumento de medida que permita establecer puntos de corte con alguna finalidad práctica o profesional. Los desarrollos estadísticos y técnicos en este campo son notables, incorporándose cada vez más a menudo los métodos estadísticos robustos (Erceg-Hurn y Mirosevich, 2008), el análisis factorial confirmatorio (Brown, 2015), los test adaptativos informatizados (Olea, Abad y Barrada, 2010; Wells y Faulkner-Bond, 2016) o el análisis de redes (Borsboom y Cramer, 2013).

10. VERSIÓN FINAL DEL TEST

En último lugar, se procede a la elaboración de la versión definitiva del test, se envía un informe de resultados a las partes interesadas y se elabora el manual que permita su utilización a otras personas o instituciones interesadas. El manual de la prueba debe recoger con todo detalle todas las características relevantes de la prueba. Finalmente, y aunque sea la última fase, esto no quiere decir que el proceso de validación concluya aquí, pues posteriores estudios deberán seguir recogiendo evidencias de validez que permitan tomar decisiones fundadas a partir de las puntuaciones de las personas. Asimismo conviene llevar a cabo una evaluación rigurosa y sistemática del instrumento elaborado, para lo cual puede utilizarse el *modelo de evaluación de test* elaborado por la European Federation of Professional Psychologists Associations (EFPA), adaptado en España por Hernández, Ponsoda, Muñiz, Prieto y Elosua (2016), que se presenta en el siguiente capítulo.

Se han descrito los diez pasos fundamentales que habría que seguir para desarrollar un test objetivo y riguroso para evaluar variables psicológicas. Estos pasos no se pueden abordar en profundidad desde un punto de vista técnico en un breve documento como este; no se trataba de eso, sino de poner a disposición de los estudiantes y profesionales una guía general que les permitiese obtener una visión panorámica de las actividades implicadas en el desarrollo de los instrumentos de medida. Se cita además la bibliografía especializada, a la que pueden acudir aquellos interesados en profundizar en esta temática. El campo de la elaboración de instrumentos de medida está altamente desarrollado, y es necesario acudir a personal cualificado para su desarrollo adecuado, pues constituye una temeridad dejarlo en manos de aficionados bienintencionados. Que un instrumento de evaluación esté adecuadamente construido y reúna las propiedades técnicas idóneas es condición necesaria, pero no es suficiente: además hay que utilizar la prueba de forma pertinente. En el capítulo 9 nos ocupamos de la utilización adecuada de los test.

9 Utilización de los test

1. ESTRATEGIAS PARA MEJORAR EL USO DE LOS TEST

En los últimos años se ha hecho un esfuerzo importante por parte de distintas instituciones y organismos para mejorar el uso de los test, y es que de nada vale que una prueba reúna las mejores características psicométricas si luego se falla a la hora de su utilización. Las organizaciones que dedican sus esfuerzos a mejorar el uso de los test, tanto nacionales como internacionales, llevan a cabo acciones y proyectos de carácter muy diverso, si bien pueden articularse en torno a dos grandes estrategias: *restrictiva* e *informativa.* La estrategia *restrictiva* se refiere a las acciones llevadas a cabo para limitar o restringir el uso de los test a aquellos profesionales que están realmente preparados para hacerlo, aunque los sistemas utilizados varían de unos países a otros (Bartram, 1996; Bartram y Coyne, 1998; Muñiz, Prieto, Almeida y Bartram, 1999; Muñiz et al., 2001; Prieto y Muñiz, 2000). Puede tratarse de restricciones legales para la comercialización de los test, certificación de profesionales que pueden usar las pruebas, restricción de acceso a los test si no se acredita la competencia como usuario, etc. Estas restricciones u otras son necesarias, pero no garantizan por sí solas un uso adecuado de los test (Moreland, Eyde, Robertson, Primoff y Most, 1995; Simner, 1996), por lo que hay que complementarlas con la difusión de información a todas las partes implicadas, tales como profesionales, usuarios, instituciones y sociedad en general. Estas acciones llevadas a cabo en el marco de la estrategia que hemos denominado «informativa» se refieren a todo tipo de iniciativas encaminadas a difundir información sobre los test y su práctica. En este sentido, distintas organizaciones nacionales e internacionales han desarrollado códigos éticos y deontológicos, así como directrices varias para guiar el uso adecuado de los test. Entre los primeros cabe destacar el metacódigo ético de la Federación Europea de Asociaciones de Psicólogos (EFPA, 1996), el código ético desarrollado por la APA (2017), el código sobre uso de los test (Joint Committee on Testing Practices, 2004) o las directrices de la Asociación Europea de Evaluación Psicológica (Fernández-Ballesteros et al., 2003). Aparte de estos códigos generales, disponemos en la actualidad de un conjunto de directrices que marcan los pasos a seguir desde la propia construcción de la prueba, su aplicación, interpretación y aplicación de los resultados (Bartram, 1998; Brennan, 2006; Downing y Haladyna, 2006; Muñiz, 1997b). Merecen mención especial los estándares técnicos desarrollados por la APA y otras dos organizaciones (AERA, APA y NCME, 2014), así como las directrices elaboradas por la Comisión Internacional de Test (ITC) para la traducción y adaptación de los test de unas culturas a otras (Hambleton, Merenda y Spielberger, 2005; International Test Commission, 2017; Muñiz, Elosua, Padilla y Hambleton, 2016). Para consultar otras directrices sobre el uso de los test en general, de los test informatizados e internet, o la utilización de los test en el ámbito del trabajo y las organizaciones, véanse, por ejemplo, los trabajos de Muñiz y Bartram (2007) y Muñiz, Hernández y Ponsoda (2015) o las páginas web de la ITC (www.intestcom.org) y de la EFPA (www.efpa.eu). También en la página web del Consejo General de Psicología de España, en el apartado de la Comisión de Test, se puede consultar información de interés (www.cop.es). Al

lado de los códigos éticos y las directrices, hay dos medidas que merecen atención dentro de las acciones enmarcadas en la estrategia de la información. Se trata por un lado de una nueva norma ISO 10667, que regula todo lo relativo a la evaluación de personas en contextos laborales, y, por otro, de los modelos de evaluación de la calidad de los test desarrollados en distintos países (Evers et al., 2013; Hernández, Ponsoda, Muñiz, Prieto y Elosua, 2016). Se presentan a continuación algunos de los aspectos más relevantes a la hora de mejorar el uso que se hace de los test.

2. FORMACIÓN DE LOS USUARIOS

¿Qué debe saber un psicólogo para utilizar los test adecuadamente? ¿Es suficiente la formación que recibe en el grado? ¿Se requiere una formación especial para el uso de ciertos test? ¿Pueden utilizar los test de forma adecuada otros profesionales que no sean psicólogos? Esas son algunas de las preguntas en torno a las cuales gira el problema de la formación. En primer lugar, señalar que el mero grado en psicología no es garantía de que se posean conocimientos suficientes para utilizar cualquier tipo de test con cualquier finalidad. En muchos países, no es el caso de España, el grado no implica pasar obligatoriamente cursos de psicometría para ser psicólogo, pero, incluso cuando es así, es imposible dar en la carrera una formación que cubra todos los test y finalidades para las que se usan. Es, por tanto, inevitable una formación continua de posgrado, que puede venir vía másteres, doctorado o cursos de especialización impartidos por distintas instituciones, tales como universidades, hospitales, organizaciones profesionales, etc. Debido a esta variedad de instituciones formadoras, existe también una amplia gama de currículums impartidos, pues la diversidad de conocimientos exigidos en función del tipo de test y del área de aplicación es muy extensa. Moreland et al. (1995), en un interesante trabajo sobre las cualificaciones de los usuarios de los test, señalan doce competencias mínimas que debe poseer cualquier usuario (tabla 9.1); pero son solo eso, los mínimos. Tal vez convenga subrayar, por obvia, la octava, relativa a la fotocopia de los materiales. Muchos psicólogos profesionales, así como profesores universitarios, no parecen darse cuenta de que la fotocopia de los materiales psicotécnicos perjudica a todos, empezando por los autores, continuando por las compañías editoras de los test y siguiendo por los propios usuarios y clientes, que no están recibiendo el material adecuado, amén de dañar la propia reputación de un profesional que trabaja con materiales fotocopiados.

TABLA 9.1

Competencias mínimas para el uso de los test

1. Evitar errores al puntuar y registrar los resultados.
2. Abstenerse de etiquetar a las personas con términos despectivos basándose en las puntuaciones de los test.
3. Mantener la seguridad de las plantillas y resto de materiales.
4. Asegurarse de que todas las personas evaluadas siguen las instrucciones.
5. Aplicar los test en unas condiciones que permitan a los evaluados un rendimiento óptimo.
6. Abstenerse de entrenar a las personas en los ítems del test.
7. Estar dispuesto a interpretar las puntuaciones y aconsejar a las personas evaluadas en sesiones diseñadas para ello.
8. No hacer fotocopias del material psicotécnico.
9. Abstenerse de utilizar hojas de respuesta caseras que pueden no ajustarse con precisión a la plantilla.
10. Establecer una buena relación con las personas evaluadas.
11. Abstenerse de responder preguntas de las personas evaluadas examinados con mayor detalle del permitido por el manual del test.
12. No asumir que una norma para un trabajo vale sin más para otro diferente y que las normas válidas para un grupo son automáticamente aplicables a otro distinto.

FUENTE: adaptado de Moreland, Eyde, Robertson, Primoff y Most (1995).

Las competencias mínimas están bien, pero los problemas sobre cuáles son los currículums más adecuados y quién debe impartirlos resultan más complejos. Por ejemplo, la comisión de test de la EFPA en una de sus reuniones (Muñiz, 1996a) discutió un posible modelo de currículum para la formación en el que se contemplasen tres factores: la

especialización requerida por el instrumento (tres niveles: *A*, *B*, *C*), el tipo de profesional (psicólogo, médicos/educadores, otros) y el área de aplicación (clínica, educativa, trabajo). Cruzando los tres factores, se tendría un modelo de 27 (3 × 3 × 3) currículums distintos, y no es exhaustivo, otros muchos modelos son pensables. En Europa cabe destacar el modelo de formación en el uso de los test en el campo de psicología del trabajo del colegio de psicólogos inglés (Bartram, 1996), así como la aproximación holandesa (Evers, 1996). Una buena panorámica de lo que ocurre en Estados Unidos puede consultarse en Fremer (1996). En nuestro país el Colegio Oficial de Psicólogos (COP) ha creado una comisión de test que está trabajando sobre este y otros asuntos relacionados con ellos. También tiene establecido un programa de formación continua (FOCAD), que incluye algunos módulos relativos a la construcción y uso de los instrumentos de medida. En suma, dada la rápida evolución del campo de la evaluación psicométrica, la formación continua resulta imprescindible para usar de forma rigurosa y responsable los instrumentos de medida.

3. ESTÁNDARES TÉCNICOS

Ahora bien, sea cual fuere el programa de formación, las normas fundamentales a las que tratan de ajustarse los formadores de usuarios son los estándares para el uso de los test, elaborados a modo de códigos técnicos por las asociaciones de psicólogos de los distintos países y entre los que destacan por su uso generalizado los editados conjuntamente por la American Educational Research Association, la APA y el National Council on Measurement in Education (2014). La primera edición de estos estándares data de 1954, con ediciones en 1955, 1966, 1974, 1985, 1999 y 2014. Los estándares cubren todos los aspectos relativos a la construcción y uso de los test, y la última edición de 2014 se organiza en torno a tres grandes apartados: fundamentos, operaciones y aplicaciones. En el apartado de fundamentos se recoge todo lo relativo a la validez, fiabilidad y equidad de los test. En el dedicado a las operaciones se aborda todo lo concerniente al diseño y construcción de los test, puntuaciones, equiparación, baremos, puntos de corte, aplicación, documentación y derechos y deberes de las personas evaluadas y de los responsables de la evaluación. Finalmente, en el apartado de aplicaciones se aborda el uso de los test en contextos aplicados, con especial atención al ámbito laboral, educativo y evaluación de programas. Estos estándares representan de algún modo el consenso psicométrico del momento, por lo que su uso y consulta resultan imprescindibles tanto en el ámbito teórico como en el aplicado y profesional.

Mención especial merece también la norma ISO 10667, que regula todo lo relativo a la evaluación de personas en entornos laborales. Las siglas ISO se refieren a la organización internacional para la estandarización (www.iso.org), que desarrolla normativas en todos sectores industriales y de servicios. En cada país tiene un representante oficial, que en el caso de España es AENOR. A iniciativa de los representantes alemanes (DIN), se inició un proceso para elaborar una nueva norma ISO que regulase todo lo relativo a la evaluación de las personas en el ámbito laboral, y tras cuatro años de trabajo de una comisión internacional la norma se publicó en 2011, y la versión española en 2013. Como es fácil de entender, esta nueva norma es de gran interés para los psicólogos, dado su papel central en la evaluación de personas en contextos laborales. La norma no inventa nada nuevo, sencillamente sistematiza y da estructura de norma ISO a las directrices, códigos éticos y regulaciones dispersas ya existentes en el ámbito de la evaluación. La norma ISO tiene una gran importancia, pues, una vez aprobada, las empresas e instituciones podrán certificarse en ella, garantizando que la cumplen. No tiene rango legal en sentido estricto pero constituye una importante base reguladora del mercado, pues no será lo mismo estar certificado que no estarlo. El objetivo de la norma es proporcionar unas reglas claras y concisas a los proveedores de servicios de evaluación y a los clientes de estos, con el fin de llevar a cabo un proceso evaluativo riguroso. Cubre todo el proceso de evaluación, desde el establecimiento del contrato de evaluación hasta la utilización de los resultados, pasando por la metodología de la evaluación en sí misma. Es aplicable a los procedimientos y métodos utilizados a nivel individual (selección, consejo, formación...), grupal (clima y cohesión de equipos de trabajo) y organizacional (clima laboral, cultura de empresa, satisfacción...). En la norma se descri-

ben las competencias, obligaciones y responsabilidades de los clientes y de los proveedores del servicio de evaluación, antes, durante y después del proceso evaluativo. También proporciona directrices para todas las partes implicadas en el proceso evaluador, incluida la propia persona evaluada y quienes reciban los resultados de la evaluación. Esta nueva norma puede suponer un importante paso para la buena práctica de la evaluación de personas en contextos laborales y organizacionales (Muñiz, 2012). Además, aporta unas ventajas claras para los psicólogos que trabajan en contextos organizacionales, pues:

a) Proporciona a los psicólogos un lenguaje claro y riguroso sobre evaluación.
b) Aporta un protocolo internacional.
c) Otorga un papel central al psicólogo en el departamento de recursos humanos.
d) Potencia el rol del departamento de recursos humanos en el organigrama de la empresa.

La norma consta de seis partes:

a) *Objeto y ámbito de aplicación,* donde se hace una introducción y una descripción general de la norma.
b) *Terminología y definiciones,* que fijan el sentido estricto de la terminología utilizada.
c) *Acuerdo cliente-proveedor de servicio,* donde se establecen las responsabilidades del cliente y del proveedor de servicios, se hace una previsión de resultados y consecuencias, se deja clara la competencia de los profesionales implicados en la evaluación y se fijan las posibles actividades complementarias de investigación a llevar a cabo.
d) *Procedimientos de preevaluación,* que tratan de identificar las necesidades evaluativas, tomar decisiones sobre los servicios de evaluación ofertados y acuerdo entre el proveedor de servicios y el cliente.
e) *Realización de la evaluación,* que constituye el núcleo de la norma y abarca los siguientes procesos: planificación de la evaluación, información a los participantes, desarrollo de la evaluación, interpretación de los resultados, preparación de informes, información a las partes implicadas *(feedback)* y valoración de la evaluación.
f) *Revisión postevaluación,* donde se comprueba el cumplimiento de los objetivos, lo que se desarrolló y no de acuerdo con lo planeado, las lecciones aprendidas para la mejora futura, las consecuencias deseadas y no deseadas que se hallaron, la claridad de los informes y el uso que se hace de ellos.

Además de estos apartados, la norma incluye cuatro anexos con especificaciones sobre:

a) Derechos y deberes de los participantes en la evaluación.
b) Documentación técnica sobre los métodos y procedimientos utilizados, que incluye todo lo relativo a la documentación y propiedades psicométricas de los instrumentos de medida utilizados (fiabilidad, validez, equidad, acomodaciones, etc.).
c) Información complementaria sobre el análisis e interpretación de los resultados.
d) Información complementaria sobre los informes y sus características.

En conclusión, la norma ISO 10667 regula con precisión y rigor las evaluaciones en entornos laborales, asegurándose de que todas las decisiones tomadas sobre las personas se basen en evidencias empíricas comprobadas.

4. PREPARACIÓN PARA LOS TEST

Cuando los test se aplican en un contexto de selección de personal, bien sea profesional o educativa, en el que la persona evaluada se juega mucho, surge inevitablemente la búsqueda de preparación o entrenamiento para intentar superar el test. En España, dado que la selectividad universitaria no se lleva a cabo mediante pruebas psicométricas, como ocurre en la mayoría de los países avanzados, este problema tiene una menor incidencia, pero es habitual en los exámenes PIR y MIR, por ejemplo, aparte de en otros procesos de selección. Cualquier área en la que el test desempeñe un papel importante para la persona evaluada es susceptible de preparación, incluso los test de integridad en el puesto de

trabajo (Alliger et al., 1996). Los datos muestran con bastante claridad (Messick y Jungeblut, 1981; Powers, 1993) que, al menos en el campo del rendimiento educativo, la preparación específica para los test logra algunas mejoras en las puntuaciones obtenidas. Las ganancias suelen ser algo mayores en las áreas cuantitativas que en las verbales, y ambas están en función, como no podía ser de otro modo, de las horas dedicadas al entrenamiento. El asunto alcanza tal envergadura económica que existen empresas transnacionales especializadas en la preparación. En Israel, por ejemplo, el 77% de los estudiantes se someten a cursos de preparación de las pruebas de selectividad universitaria (Allalouf y Shakhar, 1998). Aunque haya muchos programas distintos y modalidades, la preparación para los test siempre conlleva al menos tres elementos:

a) Familiarizarse con el test: conocer las instrucciones, tipos de ítems, tiempos, formatos, tipo de corrección, etc.
b) Revisión y estudio de los contenidos sobre los que versa el test.
c) Coger oficio *(test wiseness)* para responder al test, es decir, aprender a utilizar en beneficio propio las características y formato del test.

Las cuatro estrategias clásicas para sacar beneficio de las características del test serían: uso eficiente del tiempo disponible, evitación de errores, control sobre los aciertos al azar y razonamiento deductivo (Millman et al., 1965; Rogers y Yang, 1996).

El uso de entrenamientos para superar los test ¿puede llegar a alterar la validez predictiva de estos? Si así fuera, estaríamos ante un problema ciertamente serio. Los datos empíricos de los que se dispone (Allalouf y Shakhar, 1998; Baydar, 1990; Jones, 1986; Powers, 1985) parecen indicar con bastante claridad que al menos la validez predictiva de los test de rendimiento académico no viene negativamente afectada por la preparación. Una buena revisión de los resultados y problemas implicados en el entrenamiento para los test puede verse en los trabajos de Crocker (2006) y Bishop y Davis-Becker (2016); también los estándares de la AERA, APA y NCME (2014) incluyen algunas directrices al respecto. Emerge, no obstante, un problema adjunto, el de qué hacer con las personas que por razones económicas no pueden acceder a cursos de preparación, que no suelen ser baratos. Para mitigarlo, las instituciones encargadas de construir y aplicar la prueba deben poner en manos de los aspirantes materiales suficientes para que puedan familiarizarse con ella.

En estas situaciones de tanta relevancia para las personas evaluadas no faltará quienes intenten utilizar cualquier medio a su alcance para superar la prueba, copiando de otros, por ejemplo. Las agencias responsables de este tipo de pruebas disponen de sistemas sofisticados para detectar a los copiadores y, aparte de la meticulosidad de todo el proceso, tras la prueba llevan a cabo un escrutinio exhaustivo de las respuestas para detectar posibles anomalías. Existen varios índices estadísticos a tal efecto (Frary et al., 1977; Hanson el al., 1987), así como software específicamente diseñado para la detección a posteriori de copiadores. Un análisis detallado sobre los fraudes en el uso de las pruebas y la seguridad puede consultarse en Foster (2016), Impara y Foster (2006) o Wollack y Fremer (2013). También pueden verse las directrices de la ITC sobre seguridad (Muñiz, Hernández y Ponsoda, 2015).

5. UTILIZACIÓN DE LOS DATOS DE LOS TEST

Si todo se ha hecho bien, es decir, el test es fiable y válido y la aplicación y corrección se han llevado a cabo sin errores, aún hay que salvar el último escollo que puede hacer peligrar un uso ético y deontológico del test: la correcta utilización de los resultados. Como ya se ha visto en los códigos y estándares expuestos, este es un aspecto central en la práctica. La regla general es que los resultados solo pueden ofrecerse al cliente o persona/institución autorizada, salvo casos especiales previstos por la ley, tales como evaluaciones obligatorias. El psicólogo se mueve a veces en un terreno resbaladizo donde convergen la ética profesional, los intereses de compañías e instituciones y la legislación correspondiente. No hay reglas universales, pero el psicólogo tiene que hacer todo lo posible para que prevalezca su ética profesional. Debido a estas dificultades, la APA (1996) ha publicado un documento complementario del código ético y estándares técnicos para orientar a los psicólogos en el uso de

los datos de los test. En dicho documento se tratan problemas tales como necesidad de informar de forma exhaustiva al cliente, seguridad de los datos, personas no cualificadas para el uso de los resultados, protección del *copyright* de la prueba, peritaciones judiciales, reproducción de ítems en medios de comunicación, uso de materiales para la enseñanza y entrenamiento, etc. El denominador común a todas las recomendaciones ofrecidas es la necesidad de mantener suma prudencia a la hora de manejar tanto los resultados como los materiales del test, pues, de lo contrario, podría echarse a perder la prueba, independientemente de sus otras propiedades psicométricas.

Dentro de este contexto del uso de los datos y materiales del test, requiere mención especial la obligación legal existente en algunos países de hacer público el test *(disclosure)* una vez que se ha calificado. Los fines perseguidos con ello son nobles y pueden sintetizarse en cuatro:

a) Como cualquier otra industria, la aplicación de test debe estar sujeta al escrutinio externo que asegure transparencia y calidad.
b) Ofrece a las personas evaluadas información sobre su actuación, permitiéndoles aprender de sus errores.
c) Permite a los examinados comprobar que no se cometieron errores con ellos y que los ítems son adecuados.
d) Hace posible que las personas evaluadas se familiaricen con la prueba: formato, tipos de contenidos, tiempos, etc.

En suma, al hacer público el test, las personas que se vayan a evaluar en el futuro pueden practicar con protocolos reales del test y las ya evaluadas reciben información sobre lo hecho. Es difícil no estar de acuerdo con estos fines, pero en la práctica no siempre y en todo caso son estrictamente aplicables. La primera consecuencia de hacer público el test es que los ítems solo son utilizables una vez, lo cual encarece notablemente la prueba, pues hay que hacer nuevos ítems para cada ocasión. Por otra parte, esta filosofía encaja bien para el caso de exámenes selectivos que se realizan periódicamente, sobre todo en el campo educativo, pero choca si se trata de llevar a la práctica en áreas como la evaluación clínica, donde, por otra parte, las pruebas tienen su correspondiente *copyright.* Un problema especialmente grave se presenta si se pretende aplicar el principio sin alguna corrección a los test adaptativos informatizados. Estos test, como ya se ha señalado antes, operan aplicando ciertos ítems extraídos de un banco de ítems a cada persona evaluada. Tras examinar a unas cuantas personas, probablemente la mayoría (o todos) de los ítems del banco habrán sido utilizados, de modo que si se les obliga a publicarlos, sencillamente se quedan sin banco de ítems. Distintas organizaciones están tratando de modificar la legislación existente en algunos países, elaborada para test de papel y lápiz, para ajustarla a las nuevas exigencias de los test adaptativos informatizados (Lunz, 1997). Una posible alternativa sería, por ejemplo, publicar cada cierto tiempo un porcentaje de los ítems del banco. Aparte de aumentar los costes y chocar con los derechos de *copyright,* hacer públicos sistemáticamente los ítems de los test utilizados plantea interrogantes acerca de la influencia sobre la validez de la prueba en el futuro. Indirectamente puede tender a rebajar la calidad de la prueba, al obligar a los constructores a elaborar nuevos ítems cada vez que los utilizan una vez. Escribir ítems y contrastar sus propiedades psicométricas es tarea compleja técnicamente y nada barata en tiempo y economía. Parece que lo prudente sería encontrar un punto intermedio de equilibrio entre la necesidad innegable de que las personas evaluadas se familiaricen con las pruebas, lo cual se puede hacer de muchas formas, y el derecho de autores y editores a elaborar sus pruebas en un clima de confianza, de acuerdo con las exigencias deontológicas de la profesión.

La International Test Commission (ITC) ha publicado tres directrices de gran interés que regulan todo lo relativo a la seguridad, control de calidad y uso de los test en investigación (Muñiz, Hernández y Ponsoda, 2015). Su traducción al español de libre acceso puede verse en la página web de la Comisión de Test del Colegio Oficial de Psicólogos: www.cop.es.

6. MODELO DE EVALUACIÓN DE LA CALIDAD DE LOS TEST

Como se ha señalado más arriba, las acciones llevadas a cabo para mejorar el uso de los test pue-

den englobarse en dos grandes estrategias: restrictiva e informativa. Dentro de la estrategia informativa, una de las acciones más importante es evaluar la calidad de los test y poner estas evaluaciones a disposición de los usuarios. Con el fin de evaluar de forma rigurosa y sistemática la calidad de los test y no dejarla al criterio subjetivo de los evaluadores, la Comisión Europea de Test establecida por la EFPA ha desarrollado un modelo sistemático y estructurado de evaluación, cuya versión española se presenta más adelante. Para una descripción y análisis detallado del modelo, véase el trabajo de Hernández, Ponsoda, Muñiz, Prieto y Elosua (2016). En España el modelo se conoce por las siglas CET-R, provenientes de Cuestionario de Evaluación de Test-Revisado, puesto que existía un modelo previo desarrollado por Prieto y Muñiz (2000), que se revisó teniendo en cuenta el modelo europeo.

El objetivo del modelo no es otro que llevar a cabo una evaluación exhaustiva de la calidad de los test. Consta de tres grandes apartados. En el primero se realiza una descripción general del test en el segundo se valoran las características del test (la calidad de sus materiales, instrucciones, adaptación, desarrollo, sus ítems, etc.) y sus propiedades psicométricas (análisis de ítems, validez, fiabilidad e interpretación de las puntuaciones), y en el tercero se lleva a cabo una valoración global del test.

Una vez que se dispone de un modelo para evaluar la calidad de los test, la pregunta que surge es: ¿quién puede utilizarlo? Por supuesto, la respuesta es quien lo desee; de hecho el modelo es de libre acceso en la página web del COP. Ahora bien, en España, como ocurre en otros países, el COP ha establecido una Comisión Nacional de Test que lleva a cabo evaluaciones anuales de test y publica los resultados en su página web para que puedan consultarlas todas las personas interesadas. Para proceder a esta evaluación la Comisión de Test actúa de la siguiente manera. En primer lugar, se decide qué test se van a evaluar ese año, normalmente entre diez y doce, y luego se elige a la persona que va a coordinar la evaluación. El coordinador será el responsable de todo el proceso evaluativo, siendo su labor independiente de la Comisión de Test y de los autores y editores de estos. Este coordinador elige dos expertos que van a evaluar cada test utilizando el modelo CET-R. Si fuese necesario, puede elegir un tercer revisor. Una vez que recibe los informes de los expertos, los funde en uno solo y genera un primer informe sobre la prueba. Este informe del coordinador se envía a los editores de la prueba evaluada por si desean hacer algún comentario al respecto. Recibidos los comentarios de los editores, el coordinador tiene en cuenta lo que considere oportuno según su criterio y genera el informe definitivo, que será publicado en la página web del COP. Como se puede observar, se trata de una evaluación por pares, similar a la que se lleva a cabo para evaluar otro tipo de documentos, tales como artículos científicos, proyectos o becas. No es perfecto, nada humano lo es, pero no es el fácil encontrar otro mejor. Por ese sistema desde 2010 ya se han llevado a cabo en España cinco evaluaciones cuyos resultados pueden verse en la página web del COP y en los artículos publicados por los coordinadores de las evaluaciones (Elosua y Geisinger, 2016; Fonseca y Muñiz, 2017; Hernández, Tomás, Ferreres y Lloret, 2015; Muñiz et al., 2011; Ponsoda y Hontangas, 2013). La idea es ir avanzando hasta que se hayan evaluado la mayor parte de los test editados en España; lo ideal sería que todos los test estuviesen evaluados, como ocurre, por ejemplo, en Holanda.

La utilización de este modelo de evaluación tiene dos grandes ventajas; por un lado, ofrece información detallada de los test evaluados por expertos, lo cual es de gran ayuda para los usuarios, que disponen así de información exhaustiva de calidad sobre los test. Por otro, no menos importante, el modelo constituye una guía para editores y autores de test de cuáles son los estándares que se espera tengan sus test, contribuyendo así a la mejora de los nuevos test que se construyan o a las nuevas validaciones de los test ya existentes.

A continuación se presenta el modelo CET-R; primero se incluyen unos comentarios que se facilitan a los evaluadores antes de que apliquen el modelo y luego se presenta el modelo. Como se puede observar, el modelo es de carácter cuantitativo, pues cada característica se puntúa en una escala de 1 a 5, y también cualitativo, dado que el evaluador tiene que llevar a cabo una valoración de carácter narrativo. Esto contrasta con el modelo de evaluación americano utilizado por el instituto BUROS (Buckendahl y Plake, 2006), cuya valoración no es cuantitativa. Para un análisis de ambos modelos véase Elosua y Geisinger (2016).

Observaciones a tener en cuenta por los expertos al responder al CTR-R

1. El CET-R es un cuestionario y, como tal, sus preguntas y opciones no deben modificarse. En algunas ocasiones, al utilizar el CET original en los procesos de revisión, el revisor, al no encontrar la opción de respuesta que estaba buscando, ha modificado alguna de las opciones existentes. Esto debe evitarse. Debe, por tanto, responder valiéndose de las opciones que el CET-R ofrece. Si en alguna ocasión no encuentra la opción que está buscando, debe elegir la más similar. En las secciones para comentarios abiertos podrá hacer las aclaraciones que estime pertinentes.
2. Número de CET-R a rellenar. Si se ha de revisar una batería o más de un test (por ejemplo, el test normal y su versión abreviada), se pueden seguir dos estrategias. La que indica el CET-R es rellenar tantos CET como test haya que revisar. La que resulta menos costosa y también posible, si tiene sentido, sería utilizar solo un CET-R, dejando constancia donde corresponda de los diferentes resultados obtenidos por las distintas versiones del test.
3. Se debe revisar solo la documentación entregada. En los test comercializados se espera que el revisor realice su revisión a partir de la documentación que se le entrega. Ha ocurrido en el pasado alguna vez que el manual omite alguna información que el revisor puede considerar relevante para evaluar la calidad de la prueba. En ese caso, lo apropiado es que el revisor pida dicha información al coordinador, quien, a su vez, pedirá a la editorial la información requerida y verá si es posible obtenerla y en qué condiciones, para después distribuirla a los revisores.
4. Revisión detallada de todas las secciones del manual. Cabe señalar que, en ocasiones, a pesar de que ciertos análisis no aparecen en un apartado diferenciado (por ejemplo no figura un apartado explícito de «análisis de ítems»), los análisis sí están incluidos en el manual, aunque pasan desapercibidos para algunos revisores. Por ello, es importante que se revise con detalle toda la información del manual para que los revisores no marquen la opción «no se aporta información en la documentación» cuando aparezca en otras secciones.
5. Se espera que todas las calificaciones queden argumentadas en las preguntas abiertas que resumen cada apartado. El CET-R pide justificaciones solo de algunas respuestas a preguntas concretas, pero no de la mayoría. De todos modos, esta justificación es muy importante y debe quedar reflejada de alguna forma en los comentarios generales. Cuando se pide, por ejemplo, «comentarios sobre la validez en general», se espera encontrar información que justifique las puntuaciones dadas a todas las preguntas sobre validez.
6. Deben responderse todas las preguntas. No debería dejarse ninguna pregunta sin contestar, a no ser que en la pregunta explícitamente se indique que solo se debe contestar si se ha realizado cierto tipo de análisis. Si, para algún tipo especial de test, alguna pregunta del CET-R no resulta del todo apropiada y no se responde, la razón por la que dicha pregunta se ha dejado en blanco debería quedar justificada en el resumen de la sección correspondiente, donde se pueden introducir comentarios abiertos.
7. En test adaptados, los estudios de la versión original y la adaptada no tienen la misma relevancia. En los test adaptados de otros idiomas y/o culturas, un asunto relevante es qué peso dar a los estudios realizados con el test original y a los estudios realizados en/tras el proceso de adaptación. Nuestra posición en este asunto es que deben tenerse en cuenta todos los estudios aportados, si bien se debe dar más relevancia y peso en la evaluación a los que se aporten en el proceso de adaptación utilizando las poblaciones objetivo.
8. Términos psicométricos. En las revisiones anteriores, realizadas con el CET, hemos advertido que no todos los revisores asignan el mismo significado a los términos psicométricos empleados. Algunos términos

que inducen a error se comentan a continuación:

a) Cuando se pregunta en el apartado de «análisis de ítems» por su calidad, se pide una valoración de la información psicométrica que el manual ofrece de los ítems y no si, tras su lectura, nos parece que están bien o mal redactados.

b) Algo similar ocurre cuando se pregunta por validez de contenido. En realidad se quiere saber qué comprobaciones se aportan sobre si el test evalúa las partes relevantes del constructo de interés.

c) Por lo que se refiere a los análisis de sensibilidad y especificidad que permiten evaluar la capacidad diagnóstica del test, en ocasiones los resultados son presentados como diferencias entre grupos, y otras, como evidencias de la capacidad del test para predecir la pertenencia a un cierto grupo diagnóstico. Esta información, referida a capacidad diagnóstica, debe incluirse en evidencias de validez para predecir un criterio.

En la dirección http://glosarios.servidor-alicante.com/psicometria pueden encontrar un breve glosario de términos psicométricos que puede resultar útil.

CUESTIONARIO CET-R

1. Descripción general del test

(Si el test está compuesto de subtest heterogéneos en su formato y características, rellene un cuestionario para cada subtest. Cuando tenga sentido y sea factible, se podrá utilizar un solo CET-R, dejando constancia donde corresponda de los diferentes resultados obtenidos por los distintos subtest.)

1.1. Nombre del test.

1.2. Nombre del test en su versión original (si la versión española es una adaptación).

1.3. Autor/es del test original.

1.4. Autor/es de la adaptación española.

1.5. Editor del test en su versión original.

1.6. Editor de la adaptación española.

1.7. Fecha de publicación del test original.

1.8. Fecha de publicación del test en su adaptación española.

1.9. Fecha de la última revisión del test (si el test original es español) o de su adaptación española (si se trata de un test adaptado).

1.10. Clasifique el área general de la o las variables que pretende medir el test **(es posible marcar más de una opción)**.

(Identifique el área de contenido definido en la publicación. Si no hay una definición clara, debe señalarlo en el apartado «Otros», e indicar cuál es el área de contenido más adecuada según la información proporcionada en el manual.)

( ) Inteligencia
( ) Aptitudes
( ) Habilidades
( ) Psicomotricidad
( ) Neuropsicología
( ) Personalidad
( ) Motivación
( ) Actitudes
( ) Intereses
( ) Escalas de desarrollo
( ) Rendimiento académico/competencia curricular
( ) Escalas clínicas
( ) Potencial de aprendizaje
( ) Calidad de vida/bienestar
( ) Estrés/*burnout*
( ) Estilos cognitivos
( ) Otros (indique cuál: .....)

1.11. Breve descripción de la variable o variables que pretende medir el test.

(Se trata de hacer una descripción no evaluativa del test, con 200-600 palabras. La descripción debe proporcionar al lector una idea clara del test, lo que pretende medir y las escalas que lo conforman.)

1.12. Área de aplicación **(es posible marcar más de una opción)**.

(Identifique el área o áreas de aplicación definidas en la publicación. Si no hay una definición clara, debe señalarlo en el apartado «Otros» e indicar cuál es el área de aplicación más adecuada según la información proporcionada en el manual.)

( ) Psicología clínica
( ) Psicología educativa
( ) Neuropsicología
( ) Psicología forense
( ) Psicología del trabajo y las organizaciones
( ) Psicología del deporte
( ) Servicios sociales
( ) Salud general y bienestar
( ) Psicología del tráfico
( ) Otros (indique cuál: .....)

1.13. Formato de los ítems **(es posible marcar más de una opción)**.

( ) Respuesta construida
( ) Respuesta dicotómica (sí/no, verdadero/falso, etc.)
( ) Elección múltiple
( ) Respuesta graduada/tipo Likert
( ) Adjetivos bipolares
( ) Otro (indique cuál: .....)

1.14. Número de ítems.

(Si el test tiene varias escalas, indique el número de ítems de cada una.)

1.15. Soporte **(es posible marcar más de una opción)**.

( ) Administración oral
( ) Papel y lápiz
( ) Manipulativo
( ) Informatizado
( ) Otro (indique cuál: .....)

1.16. Cualificación requerida para el uso del test de acuerdo con la documentación aportada.

(Algunos países han adoptado sistemas para la clasificación de los test en distintas categorías en función de la cualificación requerida por los usuarios. Un sistema muy utilizado es el que divide los test en tres categorías: nivel A [test de rendimiento y conocimientos], nivel B [test colectivos de aptitudes e inteligencia] y nivel C [test de aplicación individual de inteligencia, personalidad y otros instrumentos complejos].)

( ) Ninguna
( ) Entrenamiento y acreditación especifica*
( ) Nivel A
( ) Nivel B
( ) Nivel C
( ) Otra (indique cuál: .....)

* Indique el nombre de la institución que lleva a cabo la acreditación:

1.17. Descripción de las poblaciones a las que el test es aplicable.

(Especifique el rango de edad, nivel educativo, etc., y si el test es aplicable en ciertas poblaciones específicas: minorías étnicas, personas discapacitadas, grupos clínicos, etc.)

1.18. Indique si existen diferentes formas del test y sus características (formas paralelas, versiones abreviadas, versiones informatizadas o impresas, versiones para diferentes poblaciones —infantil versus adultos—, etc.). En el caso de que existan versiones informatizadas, describa los requisitos inusuales del *hardware* y *software,* si los hubiere, que fueran necesarios para administrar correctamente el test (grabación de sonido, pantallas de resolución inusual, etc.).

1.19. Procedimiento de corrección **(es posible marcar más de una opción)**.

( ) Manual
( ) Hoja autocorregible
( ) Lectura óptica de la hoja de respuestas
( ) Automatizada por ordenador (existe *software* de corrección, o plataformas de corrección *on-line*)
( ) Efectuada por la empresa suministradora (las hojas de respuesta se envían a la empresa para que esta se ocupe de la corrección)
( ) Mediante expertos
( ) Otro (indique cuál: .....)

1.20. Puntuaciones.

(Describa el procedimiento para obtener las puntuaciones directas, totales o parciales, corrección de la probabilidad de responder correctamente por azar, inversión de ítems, etc.)

1.21. Escalas utilizadas **(es posible marcar más de una opción)**.

( ) Puntuaciones basadas en percentiles

  ( ) Centiles
  ( ) Quintiles
  ( ) Deciles

( ) Puntuaciones estandarizadas

  ( ) Puntuaciones típicas
  ( ) Eneatipos
  ( ) Decatipos
  ( ) T (media 50 y desviación típica 10)
  ( ) D (media 50 y desviación típica 20)
  ( ) CI de desviación [media 100 y desviación típica 15 (Wechsler) o 16 (Stanford-Binet)]

( ) Puntuaciones estandarizadas normalizadas (puntuaciones estandarizadas obtenidas bajo el supuesto de que su distribución es normal)
( ) Puntuaciones directas solamente
( ) Otras (indique cuál: .....)

1.22. Posibilidad de obtener informes automatizados.

( ) No
( ) Sí*

* En caso afirmativo, haga una breve valoración del informe automatizado en la que consten las características fundamentales, tales como tipo de informe y estructura, claridad, estilo, así como su calidad.

1.23. Tiempo estimado para la aplicación del test (instrucciones, ejemplos y respuestas a los ítems).

En aplicación individual: .....
En aplicación colectiva: .....

1.24. Documentación aportada por el editor **(es posible marcar más de una opción)**.

( ) Manual
( ) Libros o artículos complementarios
( ) Discos u otros dispositivos magnéticos
( ) Información técnica complementaria y actualizaciones
( ) Otra (indique cuál: .....)

1.25. Precio de un juego completo de la prueba (documentación, test, plantillas de corrección; en el caso de test informatizados, no se incluye el coste del *hardware*). Indique la fecha de consulta de precios.

1.26. Precio y número de ejemplares del paquete de cuadernillos (test de papel y lápiz). Indique la fecha de consulta de precios.

1.27. Precio y número de ejemplares del paquete de hojas de respuesta (test de papel y lápiz). Indique la fecha de consulta de precios.

1.28. Precio de la administración, y/o corrección y/o elaboración de informes por parte del editor. Indique la fecha de consulta de precios.

2. Valoración de las características del test

2.1. Calidad de los materiales del test (objetos, material impreso o *software*).

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (impresión y presentación de calidad, objetos bien diseñados, *software* atractivo y eficiente, etc.)

2.2. Calidad de la documentación aportada.

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (descripción muy clara y completa de las características técnicas, fundamentada en abundantes datos y referencias)

2.3. Fundamentación teórica.

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (descripción muy clara y documentada del constructo que se pretende medir y del procedimiento seguido para medir dicho constructo)

2.4. Adaptación del test (si el test ha sido traducido y adaptado para su aplicación en España).

( ) Característica no aplicable para este instrumento
( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena

***** ( ) Excelente [se describe con detalle el procedimiento de traducción/adaptación de los ítems a la cultura española, se estudia la equivalencia del constructo entre la versión original y adaptada, etc., es decir, se siguen las recomendaciones internacionales de traducción/adaptación de test (directrices de la ITC; véase Muñiz, Elosua y Hambleton, 2013)]

2.5. Desarrollo de los ítems del test.

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente [descripción detallada del proceso de generación de ítems, calidad de la redacción y adecuación de su formato según las directrices aceptadas (Haladyna, Downing y Rodríguez, 2002; Moreno, Martínez y Muñiz, 2006, 2015); aplicación piloto con análisis de ítems y descripción de los cambios realizados durante el proceso]

2.6. Calidad de las instrucciones para que quienes han de responder al test comprendan con facilidad la tarea.

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (claras y precisas, muy adecuadas para las poblaciones a las que va dirigido el test, incluyendo posibles acomodaciones a poblaciones especiales cuando el test también pueda aplicarse en este tipo de poblaciones)

2.7. Calidad de las instrucciones para la administración, puntuación e interpretación del test.

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (claras y precisas, tanto para la administración del test como para su puntuación e interpretación)

2.8. Facilidad para registrar las respuestas.

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (el procedimiento para emitir o registrar las respuestas es muy simple, por lo que se evitan los errores en la anotación)

2.9. Bibliografía del manual.

(Valore si en la elaboración del test se han tenido en cuenta las teorías más aceptadas sobre el constructo y el grado en que las referencias metodológicas aportadas son adecuadas)

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (reflejan una revisión adecuada y actualizada sobre el constructo y las referencias metodológicas que aporta son adecuadas)

2.10. Análisis de los ítems.

2.10.1. Datos sobre el análisis de los ítems.

( ) Característica no aplicable para este instrumento
( ) No se aporta información en la documentación
* ( ) Inadecuados
** ( ) Adecuados pero con algunas carencias
*** ( ) Adecuados
**** ( ) Buenos
***** ( ) Excelentes (información detallada sobre diversos estudios acerca de las características psicométricas de los ítems: dificultad o media, variabilidad, discriminación, validez, distractores, etc.)

2.11. Validez.

[Los estándares de la AERA, NCME y APA de 1999 y los últimos de 2014 (AERA, NCME, APA, 1999, 2014) han producido un cambio importante en el concepto de validez: no se valida el test, sino interpretaciones o usos concretos de sus puntuaciones. No hay distintos tipos de validez (de contenido, de constructo, referida al criterio, etc.), sino un tipo único. Se aceptan, eso sí, distintas fuentes de evidencias de validez. La importancia de recoger una u otra evidencia dependerá principalmente del uso que se vaya a hacer del test. De las distintas evidencias, las tres más relevantes son las basadas: *a*) en el contenido; *b*) en las relaciones con otras variables (con un criterio que se pretende predecir, con otro test que mida el mismo o un constructo relacionado, etc.), y *c*) en la estructura interna (como, por ejemplo, evaluando la estructura factorial). Los ítems que aparecen a continuación evalúan el grado en que las evidencias aportadas en cada caso son más o menos adecuadas. Si el manual del test usara la diferenciación clásica de distintos tipos de validez (por ejemplo, validez de constructo o validez referida a un criterio), se deberá incorporar la información al apartado correspondiente en función del tipo de análisis realizado.]

2.11.1. Evidencia basada en el contenido.

(Este aspecto es especialmente esencial en los test referidos al criterio y particularmente en los test de rendimiento académico. Emita su juicio sobre la calidad de la representación del contenido o dominio. Si en la documentación aportada aparecen las evaluaciones de los expertos, tómelas en consideración.)

2.11.1.1. Calidad de la representación del contenido o dominio:

( ) No se aporta información en la documentación
* ( ) Inadecuada

** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (en la documentación se presenta una precisa definición del dominio. Los ítems muestrean adecuadamente todas las facetas del dominio. Se aporta evidencia de la validez de contenido del test definitivo)

2.11.1.2. Consultas a expertos.

(Las cifras acerca del tamaño de las muestras empleadas y de los estadísticos que aparecerán más adelante tienen un carácter orientativo.)

( ) No se aporta información en la documentación
* ( ) No se ha consultado a expertos sobre la representación del contenido
** ( ) Se ha consultado de manera informal a un pequeño número de expertos
*** ( ) Se ha consultado a un pequeño número de expertos mediante un procedimiento sistematizado ($N < 10$)
**** ( ) Se ha consultado a un número moderado de expertos mediante un procedimiento sistematizado ($10 \leq N \leq 30$)
***** ( ) Se ha consultado a un amplio número de expertos mediante un procedimiento sistematizado ($N > 30$)

2.11.2. Evidencias basadas en la relación entre las puntuaciones del test y otras variables.

2.11.2.1. Relaciones con otras variables.

2.11.2.1.1. Diseños y/o técnicas empleados **(es posible marcar más de una opción)**.

( ) No se aporta información en la documentación
( ) Correlaciones con otros test
( ) Diferencias entre grupos
( ) Matriz multirrasgo-multimétodo
( ) Diseños experimentales o cuasiexperimentales
( ) Otros (indique cuál: .....)

2.11.2.1.2. Tamaño de las muestras.

[En caso de que una muestra tuviera alguna característica (por ejemplo, su carácter clínico) que pudiera justificar su tamaño reducido, indíquela.]

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña ($N < 200$)
** ( ) Un estudio con una muestra moderada ($200 \leq N \leq 500$) o varios estudios con muestras pequeñas ($N < 200$)
*** ( ) Un estudio con una muestra grande ($N > 500$)
**** ( ) Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
***** ( ) Varios estudios con muestras grandes

2.11.2.1.3. Procedimiento de selección de las muestras*.

( ) No se aporta información en la documentación
( ) Incidental
( ) Aleatorio, aunque las muestras no son representativas de la población objetivo
( ) Aleatorio, con muestras representativas de la población objetivo

* Describa brevemente el procedimiento de selección:

2.11.2.1.4. Calidad de los test marcadores empleados para evaluar las relaciones.

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (se justifica adecuadamente la selección de los test marcadores y sus propiedades psicométricas son satisfactorias)

2.11.2.1.5. Promedio de las correlaciones del test con otros test que midan el mismo constructo o constructos con los que se esperen relaciones altas.

(Se ofrecen puntos de corte para la evaluación de los coeficientes de correlación cuando se trata del mismo constructo. Dado que se esperan correlaciones de menor tamaño cuando se correlaciona el test con un constructo diferente, reduzca en 0,15 puntos los topes anteriores cuando haya de aplicarlos en esta situación.)

( ) No se aporta información en la documentación
* ( ) Inadecuada ($r < 0{,}35$)
** ( ) Adecuada pero con algunas carencias ($0{,}35 \leq r < 0{,}50$)
*** ( ) Adecuada ($0{,}50 \leq r < 0{,}60$)
**** ( ) Buena ($0{,}60 \leq r < 0{,}70$)
***** ( ) Excelente ($r \geq 0{,}70$)

2.11.2.1.6. Promedio de las correlaciones del test con otros test que midan constructos con los que el test no debería estar relacionado.

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (las correlaciones estimadas con muestras de tamaño adecuado son próximas a 0, no estadísticamente significativas, o, siendo significativas, los tamaños del efecto son bajos)

2.11.2.1.7. Resultados del análisis de la matriz multirrasgo-multimétodo.

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (los resultados apoyan tanto la validez convergente como discriminante)

2.11.2.1.8. Resultados de las diferencias entre grupos (pueden ser grupos naturales —por ejemplo, grupos demográficos— o experimentales).

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (se establecen hipótesis de validación claras y adecuadas, se observan diferencias significativas en el sentido esperado y se presta atención al tamaño del efecto)

2.11.2.2. Evidencias basadas en las relaciones entre las puntuaciones del test y un criterio.

2.11.2.2.1. Describa los criterios empleados y las características de las poblaciones.

2.11.2.2.2. Calidad de los criterios empleados.

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (se justifica adecuadamente la selección del criterio y, cuando se mida mediante un test, sus propiedades psicométricas son satisfactorias)

2.11.2.2.3. Atendiendo a la relación temporal entre la aplicación del test y la medida del criterio, indique el tipo de diseño **(es posible marcar más de una opción)**.

( ) Retrospectivo
( ) Concurrente
( ) Predictivo

2.11.2.2.4. Tamaño de las muestras en las evidencias basadas en las relaciones con un criterio*.

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña ($N < 100$)
** ( ) Un estudio con una muestra moderada ($100 \leq N < 200$) o varios estudios con muestras pequeñas ($N < 100$)

*** ( ) Un estudio con una muestra grande ($N \geq 200$)
**** ( ) Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
***** ( ) Varios estudios con muestras grandes o estudios metaanalíticos apropiados

* Indique si alguna característica de las muestras (por ejemplo, su carácter clínico) pudiera justificar el tamaño reducido de la o las muestras:

2.11.2.2.5. Procedimiento de selección de las muestras*.

( ) No se aporta información en la documentación
( ) Incidental
( ) Aleatorio, aunque las muestras no son representativas de la población objetivo
( ) Aleatorio, con muestras representativas de la población objetivo

* Describa brevemente el procedimiento de selección y proporcione información relevante sobre el grado de representatividad de las muestras.

2.11.2.2.6. Promedio de las correlaciones del test con los criterios.

(Los rangos de valores mostrados abajo se refieren a la correlación entre el test y un criterio, que es la forma más habitual de obtener evidencias de validez referida a un criterio. Sin embargo, en ciertas situaciones clínicas, como cuando se usan test de *screening* en un proceso diagnóstico, puede resultar más útil proporcionar información sobre la sensibilidad y especificidad del test, por ejemplo mediante curvas ROC, que correlaciones. En estos casos, el revisor deberá tener en cuenta la sensibilidad y especificidad del test a la hora de evaluar su utilidad para tomar decisiones diagnósticas y así determinar el nivel de adecuación del test, añadiendo los comentarios pertinentes en la sección 2.11.5.)

( ) No se aporta información en la documentación
* ( ) Inadecuada ($r < 0{,}20$)
** ( ) Suficiente ($0{,}20 \leq r < 0{,}35$)
*** ( ) Buena ($0{,}35 \leq r < 0{,}45$)
**** ( ) Muy buena ($0{,}45 \leq r < 0{,}55$)
***** ( ) Excelente ($r \geq 0{,}55$)

2.11.3. Evidencias basadas en la estructura interna del test.

2.11.3.1. Resultados del análisis factorial (exploratorio y/o confirmatorio).

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena

***** ( ) Excelente (los resultados apoyan la estructura del test tanto en lo que se refiere al número de factores extraídos como a su interpretación. Además, se proporciona información suficiente y adecuada para evaluar la calidad de las decisiones tomadas al aplicar la técnica —AFE y/o AFC, método de factorización, rotación, *software* empleado, etc.— e interpretar los resultados)

2.11.3.2. Datos sobre el funcionamiento diferencial de los ítems.

( ) No se aporta información en la documentación
* ( ) Inadecuados
** ( ) Adecuados pero con algunas carencias
*** ( ) Adecuados
**** ( ) Buenos
***** ( ) Excelentes (información detallada sobre diversos estudios acerca del sesgo de los ítems relacionado con el sexo, la lengua materna, etc. Empleo de la metodología apropiada)

2.11.4. Indique si el manual del test informa de las acomodaciones a introducir en la administración del test, para la correcta evaluación de personas con limitaciones o diversidad funcional.

( ) No
( ) Sí*

* En caso afirmativo, indique cuáles y si se han justificado adecuadamente en el manual.

2.11.5. Comentarios sobre la validez en general.

(Resuma, por favor, las principales evidencias de validez que la documentación examinada aporta, estime su calidad y justifique las puntuaciones otorgadas en las preguntas previas. En caso de que haya revisado información sobre la sensibilidad y especificidad del test, por ejemplo mediante curvas ROC, los resultados a la hora de determinar la utilidad diagnóstica del test serán comentados en este punto. También se comentará cualquier otro tipo de evidencia diferente de las consideradas en el modelo, por ejemplo basadas en el proceso de respuesta, si las hubiere.)

2.12. Fiabilidad

2.12.1. Datos aportados sobre la fiabilidad **(es posible marcar más de una opción)**.

( ) Un único coeficiente de fiabilidad (para cada escala o subescala)
( ) Varios coeficientes de fiabilidad (para cada escala o subescala)
( ) Un único error típico de medida (para cada escala o subescala)
( ) Coeficientes de fiabilidad para diferentes grupos de personas
( ) Error típico de medida para diferentes grupos de personas

( ) Cuantificación del error mediante TRI (función de información u otros)
( ) Otros indicadores de fiabilidad (indique cuáles: .....)

2.12.2. Equivalencia (formas paralelas).
(Rellenar solo si es aplicable al instrumento.)

2.12.2.1. Tamaño de las muestras en los estudios de equivalencia.

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña ($N < 200$)
** ( ) Un estudio con una muestra moderada ($200 \leq N < 500$) o varios estudios con muestras pequeñas ($N < 200$)
*** ( ) Un estudio con una muestra grande ($N > 500$)
**** ( ) Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
***** ( ) Varios estudios con muestras grandes

2.12.2.2. Resultados de la puesta a prueba de los supuestos de paralelismo.

( ) No se aporta información en la documentación
* ( ) Inadecuados
** ( ) Adecuados pero con algunas carencias
*** ( ) Adecuados
**** ( ) Buenos
***** ( ) Excelentes (se realizan pruebas de significación para poner a prueba la igualdad de las medias y de las varianzas de las formas, así como la igualdad de las correlaciones con otros test)

2.12.2.3. Promedio de los coeficientes de equivalencia.

( ) No se aporta información en la documentación
* ( ) Inadecuada ($r < 0{,}50$)
** ( ) Adecuada pero con algunas carencias ($0{,}50 \leq r < 0{,}60$)
*** ( ) Adecuada ($0{,}60 \leq r < 0{,}70$)
**** ( ) Buena ($0{,}70 \leq r < 0{,}80$)
***** ( ) Excelente ($r \geq 0{,}80$)

2.12.3. Consistencia interna.
(Rellenar solo si es aplicable al instrumento.)

2.12.3.1. Tamaño de las muestras en los estudios de consistencia.

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña ($N < 200$)
** ( ) Un estudio con una muestra moderada ($200 \leq N < 500$) o varios estudios con muestras pequeñas ($N < 200$)
*** ( ) Un estudio con una muestra grande ($N \geq 500$)
**** ( ) Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
***** ( ) Varios estudios con muestras grandes

2.12.3.2. Coeficientes de consistencia interna presentados.

( ) No se aporta información
( ) Coeficiente alfa o KR-20
( ) Alfa ordinal
( ) Lambda-2
( ) Otro (indique cuál: .....)

2.12.3.3. Promedio de los coeficientes de consistencia.

( ) No se aporta información en la documentación
* ( ) Inadecuada ($r < 0{,}60$)
** ( ) Adecuada pero con algunas carencias ($0{,}60 \leq r < 0{,}70$)
*** ( ) Adecuada ($0{,}70 \leq r < 0{,}80$)
**** ( ) Buena ($0{,}80 \leq r < 0{,}85$)
***** ( ) Excelente ($r \geq 0{,}85$)

2.12.4. Estabilidad (test-retest).
(Rellenar solo si es aplicable al instrumento.)

2.12.4.1. Tamaño de las muestras en los estudios de estabilidad:

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña ($N < 100$)
** ( ) Un estudio con una muestra moderada ($100 \leq N < 200$) o varios estudios con muestras pequeñas ($N < 100$)
*** ( ) Un estudio con una muestra grande ($N \geq 200$)
**** ( ) Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
***** ( ) Varios estudios con muestras grandes

2.12.4.2. Promedio de los coeficientes de estabilidad.

( ) No se aporta información en la documentación
* ( ) Inadecuada ($r < 0{,}55$)
** ( ) Adecuada pero con algunas carencias ($0{,}55 \leq r < 0{,}65$)
*** ( ) Adecuada ($0{,}65 \leq r < 0{,}75$)
**** ( ) Buena ($0{,}75 \leq r < 0{,}80$)
***** ( ) Excelente ($r \geq 0{,}80$)

2.12.5. Cuantificación de la precisión mediante TRI.
(Rellenar solo si se ha empleado TRI.)

2.12.5.1. Tamaño de las muestras en los estudios de TRI.

[Depende del formato de los ítems y del modelo empleado. Como referencia, en el caso de los modelos para datos dicotómicos, unas recomendaciones generales sobre el tamaño adecuado son 200 casos para el modelo de un parámetro, 400 para el modelo de dos parámetros y 700 para el de tres (Parshall, Spray, Kalohn y Davey, 2002).]

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña
** ( ) Un estudio con una muestra adecuada
*** ( ) Un estudio con una muestra grande
**** ( ) Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
***** ( ) Varios estudios con muestras grandes

2.12.5.2. Coeficientes proporcionados.

( ) No se aporta información
( ) Fiabilidad de las puntuaciones en el rasgo latente
( ) Función de Información
( ) Otro (indique cuál: .....)

2.12.5.3. Tamaño de los coeficientes.

[Debe tenerse en cuenta que el valor de los coeficientes depende del valor del rasgo latente, existiendo típicamente un rango de puntuaciones latentes para el que el test es óptimo en términos de precisión. Para la valoración del tamaño de los coeficientes, más que ese rango óptimo se debe tener en cuenta el rango de puntuaciones para el que los resultados del test pueden tener importancia. Si no existe tal rango a priori, la evaluación debe basarse en la información promedio proporcionada (Reise y Haviland, 2005). A continuación se proporcionan valores orientativos para la información promedio del test, si bien estos valores deben usarse con cautela por la poca experiencia existente en la aplicación de estos puntos de corte, y porque dependen del número de ítems del test.]

( ) No se aporta información en la documentación
* ( ) Inadecuada (información < 2)
** ( ) Adecuada pero con algunas carencias (2 ≤ información < 3,33)
*** ( ) Adecuada (3,33 ≤ información < 5)
**** ( ) Buena (5 ≤ información < 10)
***** ( ) Excelente (información ≥ 10)

2.12.6. Fiabilidad interjueces.

(Rellenar solo si es aplicable al instrumento.)

2.12.6.1. Coeficientes de fiabilidad inter-jueces **(es posible marcar más de una opción)**.

( ) Porcentaje de acuerdo
( ) Coeficiente *kappa*
( ) Coeficiente de correlación intraclase (ICC)
( ) Coeficiente basado en la teoría de la generalizabilidad
( ) Otro (indique cuál: .....)

2.12.6.2. Valor promedio de los coeficientes de fiabilidad interjueces.

(Se ofrecen a continuación unos puntos de corte orientativos.)

( ) No se aporta información en la documentación
* ( ) Inadecuado ($r < 0{,}50$)
** ( ) Adecuado pero con algunas carencias ($0{,}50 < r \leq 0{,}60$)

*** ( ) Adecuado ($0{,}60 < r \leq 0{,}70$)
**** ( ) Bueno ($0{,}70 < r \leq 0{,}80$)
***** ( ) Excelente ($r > 0{,}80$)

2.12.7. Comentarios sobre la fiabilidad en general.

[Resuma los resultados que ha extraído de la documentación sobre la fiabilidad de las puntuaciones. Comente los distintos tipos de indicadores obtenidos, el rango de los coeficientes, si los resultados están basados en muestras adecuadas, para qué poblaciones (en función de los resultados de TRI) resulta el test más preciso, etc. Justifique las valoraciones otorgadas a las preguntas precedentes.]

2.13. Baremos e interpretación de puntuaciones.

[A la hora de interpretar las puntuaciones, se puede diferenciar entre una interpretación normativa o una referida a un criterio. La interpretación normativa se deriva de comparar la puntuación del evaluado con la distribución de las puntuaciones observadas en un grupo de referencia. La interpretación referida a un criterio o dominio requiere el establecimiento de puntos de corte que reflejen el dominio, o no, de una serie de competencias o aptitudes o si, en escalas clínicas, la persona supera un punto de corte que refleje la necesidad de una intervención, por ejemplo. A veces los puntos de corte se establecen a partir del juicio de expertos, y otras a partir de investigaciones empíricas que permiten realizar clasificaciones y asignar a las personas a diferentes programas de intervención. Se debe responder a uno o a los dos apartados (interpretación normativa y/o interpretación referida a un criterio) en función de la interpretación de las puntuaciones considerada en el manual.]

2.13.1. Interpretación normativa.

(Responder solo si es aplicable al test.)

2.13.1.1. Calidad de las normas.

[Desde ciertas posiciones teóricas y metodologías, como la tipificación continua *(continuous norming),* la generación de un número reducido de baremos no indica necesariamente una baja calidad de la información ofrecida para la interpretación de las puntuaciones. La tipificación continua utiliza la información disponible de todos los grupos para construir el baremo de cada grupo concreto, lo que resulta en baremos más precisos con grupos más reducidos (Evers, Sijtsma, Lucassen y Meijer, 2010; Zachary y Gorsuch, 1985). Tenga en cuenta esta posibilidad a la hora de emitir su valoración en este ítem.]

( ) No se aporta información en la documentación
* ( ) Un baremo que no es aplicable a la población objetivo
** ( ) Un baremo aplicable a la población objetivo con cierta precaución, considerando las diferencias entre poblaciones
*** ( ) Un baremo adecuado para la población objetivo
**** ( ) Varios baremos dirigidos a diversos estratos poblacionales
***** ( ) Amplio rango de baremos en función de la edad, el sexo, el nivel cultural y otras características relevantes

2.13.1.2. Tamaño de las muestras.

(Si hay varios baremos, clasifique el tamaño promedio.)

( ) No se aporta información en la documentación

* ( ) Pequeño ($N < 150$)
** ( ) Suficiente ($150 \leq N < 300$)
*** ( ) Moderado ($300 \leq N < 600$)
**** ( ) Grande ($600 \leq N < 1.000$)
***** ( ) Muy grande ($N \geq 1.000$)

2.13.1.3. Indique si se ha aplicado una estrategia de tipificación continua *(continuous norming)* usando diferentes grupos de edad, para conseguir baremos de más calidad.

( ) Sí
( ) No

2.13.1.4. Procedimiento de selección de las muestras*.

( ) No se aporta información en la documentación
( ) Incidental
( ) Aleatorio, aunque las muestras no son representativas de la población objetivo
( ) Aleatorio, con muestras representativas de la población objetivo

* Describa brevemente el procedimiento de selección.

2.13.1.5. Actualización de los baremos.

( ) No se aporta información en la documentación
* ( ) Inadecuada (más de 25 años)
** ( ) Adecuada pero con algunas carencias (entre 20 y 24 años)
*** ( ) Adecuada (entre 15 y 19 años)
**** ( ) Buena (entre 10 y 14 años)
***** ( ) Excelente (menos de 10 años)

2.13.2. Interpretación referida a un criterio.
(Responder solo si es aplicable al test.)

2.13.2.1. Adecuación del establecimiento de los puntos de corte establecidos.

( ) No se aporta información en la documentación
* ( ) Inadecuado
** ( ) Adecuado pero con algunas carencias
*** ( ) Adecuado
**** ( ) Bueno
***** ( ) Excelente (se cuenta con un grupo de un mínimo de tres o cuatro jueces con formación y experiencia en el ámbito de estudio, y/o se proporciona evidencia empírica con estudios de calidad que relacionan el punto de corte con un criterio externo, para avalar la adecuación y utilidad de los puntos de corte establecidos)

2.13.2.2. Si se utiliza el juicio de expertos para establecer los puntos de corte, indique el procedimiento empleado para fijar el estándar.

( ) Nedelsky
( ) Angoff
( ) Zieky y Livingston
( ) Hofstee
( ) Otro (indique cuál: .....)

2.13.2.3. Si se utiliza el juicio de expertos para establecer los puntos de corte, indique cómo se ha obtenido el acuerdo interjueces **(es posible marcar más de una opción)**.

( ) Coeficiente ρ_0
( ) Coeficiente *kappa*
( ) Coeficiente Livingston
( ) Coeficiente de correlación intraclase (ICC)
( ) Otro (indique cuál: .....)

2.13.2.4. Si se utiliza el juicio de expertos para establecer los puntos de corte, indique el valor del coeficiente de acuerdo interjueces (por ejemplo, *kappa* o ICC).

( ) No se aporta información en la documentación
* ( ) Inadecuado ($r < 0{,}50$)
** ( ) Adecuado pero con algunas carencias ($0{,}50 < r \leq 0{,}60$)
*** ( ) Adecuado ($0{,}60 < r \leq 0{,}70$)
**** ( ) Bueno ($0{,}70 < r \leq 0{,}80$)
***** ( ) Excelente ($r > 0{,}80$)

2.13.3. Comentarios sobre los baremos y establecimientos de puntos de corte.
(Resuma y evalúe los procedimientos que el test propone para facilitar la interpretación de las puntuaciones y justifique las evaluaciones dadas a las preguntas precedentes.)

3. Valoración global del test

3.1. Con una extensión máxima de 1.000 palabras, exprese su valoración del test, resaltando sus puntos fuertes y débiles, así como recomendaciones acerca de su uso en diversas áreas profesionales. Indique asimismo cuáles son las características de la prueba que podrían ser mejoradas, carencias de información en la documentación, etc.

A modo de resumen, rellene las tablas 1 y 2.
La tabla 1 incluye algunos datos descriptivos del test.

TABLA 1
Descripción del test

Característica	Apartado	Descripción
Nombre del test	1.1	
Autor	1.3	
Autor de la adaptación española	1.4	
Fecha de la última revisión	1.9	
Constructo evaluado	1.11	
Áreas de aplicación	1.12	
Soporte	1.15	

En la tabla 2 se resume la valoración de las características generales del test. Tome en consideración el promedio de las calificaciones emitidas en los apartados que figuran en la segunda columna de la tabla 2. El número de asteriscos que acompaña a las opciones de respuesta de los ítems se corresponde con la puntuación correspondiente a cada ítem (de 1. «Inadecuado» a 5. «Excelente»)

TABLA 2
Valoración del test

Característica	Apartados	Valoración
Materiales y documentación	2.1 y 2.2	
Fundamentación teórica	2.3	
Adaptación	2.4	
Análisis de ítems	2.10	
Validez: contenido	2.11.1	
Validez: relación con otras variables	2.11.2	
Validez: estructura interna	2.11.3	
Validez: análisis del DIF	2.11.3.2	
Fiabilidad: equivalencia	2.12.2	
Fiabilidad: consistencia interna	2.12.3	
Fiabilidad: estabilidad	2.12.4	
Fiabilidad: TRI	2.12.5	
Fiabilidad interjueces	2.12.6	
Baremos e interpretación de puntuaciones	2.13	

10 Mirando hacia el futuro

Predecir el futuro es tarea imposible, el de la evaluación psicométrica incluido, pues como bien nos advirtió Taleb (2008) en su libro *El cisne negro,* nadie hasta ahora fue capaz de prever los grandes acontecimientos que a la postre cambiaron el rumbo de la humanidad. No se trata, por tanto, de predecir aquí el rumbo que tomará la evaluación psicológica del futuro lejano, sino de señalar las vías que se vislumbran, basándonos en las tendencias actuales a partir de las cuales se va desarrollando la disciplina. Nos apoyaremos para ello en trabajos previos sobre el tema (Muñiz, 2012, 2018; Muñiz y Fernández-Hermida, 2010; Muñiz, Hernández y Ponsoda, 2015). Como ya se ha señalado varias veces a lo largo del libro, la gran fuerza que está remodelando la evaluación psicológica en la actualidad son las nuevas tecnologías de la información, y en especial los avances informáticos, multimedia e internet. Autores como Bennet (1999, 2006), Breithaupt, Mills y Melican (2006), Drasgow (2016), Drasgow, Luecht y Bennet (2006) o Sireci y Faulkner-Bond (2016), entre otros muchos, consideran que las nuevas tecnologías están influyendo sobre todos los aspectos de la evaluación psicológica, tales como el diseño de los test, la construcción y presentación de los ítems, la puntuación de los test y la evaluación a distancia. Emergen nuevas formas de evaluación, aunque, no nos engañemos, los test psicométricos seguirán siendo herramientas fundamentales, dada su objetividad y economía de medios y tiempo (Phelps, 2005, 2008). En este contexto de cambio tecnológico surge la llamada psicología 2.0 (Armayones et al., 2015), que pretende extender la psicología a través de las facilidades que ofrecen internet y las redes sociales. La evaluación no puede estar ajena a estas nuevas tendencias, apareciendo nuevos enfoques psicométricos conectados con el análisis de las grandes bases de datos *(big data)* de las que se dispone actualmente (Markovetz, Blaszkiewicz, Montag, Switala y Schlaepfer, 2014). Por ejemplo, las ventajas potenciales de usar los teléfonos móviles como terminales para la evaluación abren nuevas posibilidades para la psicometría del futuro (Armayones et al., 2015; Chernyshenko y Stark, 2016; Miller, 2012). Trabajos como el pionero de Kosinski, Stillwell y Graepel (2013) analizan con éxito la posibilidad de utilizar los «me gusta» de Facebook como predictores de distintas características humanas, entre ellas los rasgos de la personalidad, lo que hace preguntarse si nuestros rastros en las redes sociales sustituirán algún día no muy lejano a los cuestionarios y test tal como los conocemos ahora. No sabemos nada del futuro, pero se nos representa bello y excitante, una lucha sorda de fondo entre nuestra inteligencia de carbono y agua y la artificial del silicio. No sabemos si una de ellas vencerá a la otra, o se producirá la simbiosis, pero lo que está claro es que el silicio reclama un mayor rol en nuestras vidas, y la evaluación psicométrica no es una excepción. Eso sí, la prueba del algodón, el árbitro, siempre será la validez; todas las fantasías y avances tecnológicos pasan por demostrar que aportan mejoras en la medida del constructo evaluado, pues de lo contrario no dejarán de ser meros fuegos de artificio.

Según Hambleton (Hambleton, 2004, 2006, 2009), seis grandes áreas están atrayendo la atención de investigadores y profesionales. La primera es el uso internacional de los test, lo que plantea todo un conjunto de problemas de adaptación de

los test de unos países a otros (Byrne et al., 2009; Hambleton et al., 2005; Muñiz et al., 2016). La segunda es el uso de nuevos modelos psicométricos y tecnologías para generar y analizar los test. La tercera es la aparición de nuevos formatos de ítems derivados de los grandes avances informáticos y multimedia, pasando de las modestas matrices en blanco y negro a las pantallas interactivas, con animación y sonido, capaces de reaccionar a las respuestas de las personas evaluadas (Irvine y Kyllonen, 2002; Shermis y Burstein, 2013; Sireci y Zenisky, 2006, 2016). La cuarta área que reclamará gran atención es todo lo relacionado con los test informatizados y sus vínculos con internet. Como ya se ha comentado, mención especial merecen en este campo los test adaptativos informatizados que permiten ajustar la prueba a las características de la persona evaluada, sin por ello perder objetividad o comparabilidad entre las personas, lo cual abre perspectivas muy prometedoras en la evaluación (Mills y Breithaupt, 2016; Zenisky y Luecht, 2016). La evaluación a distancia o teleevaluación es otra línea que se abre camino con rapidez, lo cual planta serios problemas de seguridad de los datos y de las personas, pues hay que comprobar que la persona que se está evaluando es la que realmente dice ser, sobre todo en contextos de selección de personal o de pruebas con importantes repercusiones para la vida futura de la persona evaluada. En este campo se están dando grandes avances básicos y aplicados (Bartram y Hambleton, 2006; Leeson, 2006; Mills et al., 2002; Parshall et al., 2002; Williamson et al., 2006; Wilson, 2005). En quinto lugar cabe señalar un campo que puede parecer periférico pero que está cobrando gran importancia. Se trata de los sistemas a utilizar para dar los resultados a los usuarios y partes legítimamente implicadas. Es fundamental que estos comprendan sin equívocos los resultados de las evaluaciones, y no es obvio cuál es la mejor manera de hacerlo, sobre todo si se tienen que enviar para la interpretación y explicación del profesional, como ocurre en numerosas situaciones de selección de personal o en la evaluación educativa (Goodman y Hambleton, 2004; Zenisky y Hambleton, 2016). Finalmente es muy probable que en el futuro haya una gran demanda de formación por parte de distintos profesionales relacionados con la evaluación; estar al tanto de los cambios exige formación continua.

Por su parte, Sireci y Faulkner-Bond (2016) subrayan seis tendencias actuales en línea con las ya comentadas del profesor Hambleton: uso de los test para establecer responsabilidades sobre la calidad de la educación, hacer las evaluaciones más accesibles y adaptables a todas las personas, sean cuales sean sus características personales, aumento de las evaluaciones internacionales, uso de las nuevas tecnologías para mejorar la evaluación (Drasgow, 2016), demanda de nuevos métodos para mejorar los informes de los resultados y la evaluación diagnóstica y finalmente llevar a cabo evaluaciones en contextos menos estructurados, como los juegos (gamificación), utilizando además el potencial formativo de las evaluaciones. En el reciente libro sobre tecnología y test editado por Drasgow (2016) se incluyen interesantes capítulos sobre nuevos tipos de ítems, evaluación y juegos, simulaciones, ensamblaje automático de los test, corrección automática de las pruebas, evaluación ambulatoria, entre otros, que dan una buena idea de por dónde emergen las innovaciones en nuestros días.

Otro tema que cobra pujanza es el de la evaluación ambulatoria ya citada, que si bien tiene rancio abolengo en psicología, está resurgiendo con fuerza en la actualidad impulsada por las nuevas tecnologías (Chernyshenko y Stark, 2016; Trull y Ebner-Priemer, 2009, 2013; Van Os, Delespaul, Wigman, Myin-Germeys y Wichers, 2013). La evaluación ambulatoria abarca una amplia gama de métodos de evaluación que tratan de estudiar las experiencias de las personas en su entorno natural y en la vida diaria, permitiendo evaluar determinadas variables y constructos psicológicos desde una perspectiva más dinámica, personalizada, contextual y ecológica. Permite evaluar los sentimientos, las cogniciones, las emociones y los síntomas de las personas mediante dispositivos móviles en su contexto real diario. Para ello habitualmente se realizan evaluaciones varias veces al día durante un período temporal (típicamente una semana) para captar suficientemente la variabilidad de los fenómenos. Las preguntas se activan mediante un *beep* en un marco temporal fijado por el investigador, por ejemplo, entre las diez de la mañana y las diez de la noche. Además, estos *beeps* pueden presentarse de forma aleatoria o en intervalos de tiempo predeterminados, por ejemplo cada 90 minutos. A lo largo de cada día se recogen diferentes muestras de compor-

tamiento, aproximadamente seis u ocho por día durante siete días. Todos estos datos se vuelcan a una plataforma para su análisis posterior. Se trata, pues, de un abordaje complementario a los procedimientos tradicionales de evaluación psicométrica de papel y lápiz en contextos más o menos artificiales y de corte más bien transversal y retrospectivo (Fonseca y Muñiz, 2017). La flexibilidad de los nuevos modelos psicométricos de análisis de redes pueden permitir la incorporación y análisis de este tipo de datos (Borsboom y Cramer, 2013; Fonseca, 2017), así como los modelos procedentes de la teoría de los sistemas dinámicos o la teoría del caos (Nelson, McGorry, Wichers, Wigman y Hartmann, 2017).

Otro reto fundamental al que se enfrenta la evaluación psicológica es el uso masivo de autoinformes en detrimento de otros indicadores de carácter neurobiológico, personas cercanas *(proxies)* u observación conductual, entre otros. Ahora bien, los autoinformes tienen serias limitaciones a dos niveles: epistemológico y técnico. A nivel espistemológico, al hacer que una persona informe sobre sí misma retrotraemos la psicología al estatus de ciencia introspectiva, dejando nuestro nivel de análisis al albur de lo que una persona cree saber sobre sí misma, o decida decirnos. A nivel técnico, los autoinformes resultan muy vulnerables al falseamiento y la distorsión, por lo que los hace inservibles en numerosas situaciones (Arendasy, Sommer, Herle, Schützhofer e Inwanschitz, 2011; Hogan, Barrett y Hogan, 2007). Para evitar estos inconvenientes se están desarrollado numerosas estrategias, destacando las pruebas ipsativas (Brown y Maydeu-Olivares, 2013) y los test de asociación implícita (IAT), los cuales permiten detectar la asociación automática que una persona muestra sobre diferentes ideas, objetos o conceptos (Greenwald y Banaji, 1995; Greenwald et al., 2009), evitando así la distorsión consciente de los autoinformes. Como cualquier otra tecnología emergente, los IAT no están exentos de limitaciones y existe un debate sobre ellos en la literatura especializada (Barth, 2007; Fazio y Olson, 2003; Gawronsky y Payne, 2010; Hofmann, Gawronski, Gschwendner, Le y Schmitt, 2005).

Estas son algunas líneas de trabajo y los retos sobre los que muy probablemente girarán las actividades evaluadoras en un futuro no muy lejano. No se trata de hacer una relación exhaustiva ni mucho menos, sino de indicar algunas pistas para orientarse en el mundo cambiante de la evaluación psicológica. Estos cambios y progresos que se están produciendo en la evaluación psicológica son de vital importancia, pues al fin y al cabo la evaluación rigurosa constituye la base de unos diagnósticos precisos, claves a su vez para generar intervenciones eficaces.

Apéndice

1.1. $e = X - V$

Según el modelo:

$$X = V + e$$

Despejando:

$$e = X - V$$

1.2. $E(e) = 0$

Según 1.1:

$$e = X - V$$

Esperanza matemática: $E(e) = E(X - V)$

$$E(e) = E(X) - E(V)$$

Según el modelo para una puntuación verdadera dada:

$$E(X) = V$$

luego:

$$E(e) = V - E(V) = V - V = 0$$

1.3. $\mu_x = \mu_v$

Según el modelo: $X = V + e$

$$E(X) = E(V + e) = E(V) + E(e)$$

Según 1.2:

$$E(e) = 0$$

luego

$$E(X) = E(V)$$
$$\mu_x = \mu_v$$

1.4. $\text{cov}(V, e) = 0$

La covarianza entre V y e vendrá dada según la definición de covarianza por:

$$\text{cov}(V, e) = \rho_{Ve}\sigma_V\sigma_e$$

Según el supuesto 2 del modelo:

$$\rho_{Ve} = 0$$

luego:

$$\text{cov}(V, e) = (0)\sigma_V\sigma_e = 0$$

1.5. $\text{cov}(X, V) = \text{var}(V)$

La covarianza entre X y V vendrá dada por

$$\text{cov}(X, V) = E(XV) - E(X)E(V)$$

Sustituyendo X por su valor en el modelo: $X = V + e$

$$\begin{aligned}\text{cov}(X,V) &= E[(V + e)V] - E(V + e)E(V) = \\ &= E(V^2) + E(Ve) - E(V)E(V) - \\ &- E(e)E(V)\end{aligned}$$

Ahora bien,

$$E(Ve) - E(V)E(e) = \text{cov}\,(V, e)$$

Y según 1.4:

$$\text{cov}\,(V, e) = 0$$

luego

$$\text{cov}\,(X, V) = E(V^2) - [E(V)]^2 = \text{var}\,(V),$$

ya que

$$E(V^2) - [E(V)]^2 = \text{var}\,(V)$$

1.6. $\text{cov}\,(X_j, X_k) = \text{cov}\,(V_j, V_k)$

La covarianza entre X_j y X_k vendrá dada por:

$$\text{cov}\,(X_j, X_k) = E(X_j, X_k) - E(X_j)E(X_k)$$

Sustituyendo X_j y X_k por su valor según el modelo:

$$\begin{aligned}\text{cov}\,(X_j, X_k) = {} & E[(V_j + e_j)(V_k + e_k)] - \\ & - E(V_j + e_j)E(V_k + e_k) = \\ & = E(V_jV_k) + E(V_je_k) + \\ & + E(e_jV_k) + E(e_je_k) - \\ & - E(V_j)E(V_k) - E(V_j)E(e_k) - \\ & - E(e_j)E(V_k) + E(e_j)E(e_k)\end{aligned}$$

Ahora bien:

$$\begin{aligned}E(V_je_k) - E(V_j)E(e_k) &= \text{cov}\,(V_j, e_k) \\ E(e_jV_k) - E(e_j)E(V_k) &= \text{cov}\,(e_j, V_k) \\ E(e_je_k) - E(e_j)E(e_k) &= \text{cov}\,(e_j, e_k)\end{aligned}$$

Y según los supuestos 2 y 3 del modelo es inmediato que:

$$\begin{aligned}\text{cov}\,(V_j, e_k) &= 0 \\ \text{cov}\,(e_j, V_k) &= 0 \\ \text{cov}\,(e_j, e_k) &= 0\end{aligned}$$

dado que las puntuaciones verdaderas no covarían con los errores ni los errores entre sí. Luego nos queda que:

$$\begin{aligned}\text{cov}\,(X_j, X_k) &= (V_jV_k) - E(V_j)E(V_k) = \\ &= \text{cov}\,(V_j, V_k)\end{aligned}$$

Nota. Si se tratase de *formas paralelas,* entonces V_j y V_k serían iguales; por tanto, su covarianza sería la varianza, pudiendo escribirse [1.6]:

$$\text{cov}\,(X_j, X_k) = \text{cov}\,(V_j, V_k) = \text{var}\,(V)$$

1.7. $\text{var}\,(X) = \text{var}\,(V) + \text{var}\,(e)$.

Según el modelo:

$$X = V + e$$

La varianza de una variable compuesta viene dada por:

$$\text{var}\,(X) = \text{var}\,(V) + \text{var}\,(e) + 2\,\text{cov}\,(V, e)$$

Ahora bien, según 1.4,

$$\text{cov}\,(V, e) = 0$$

luego

$$\text{var}\,(X) = \text{var}\,(V) + \text{var}\,(e)$$

1.8. $\rho_{xe} = \sigma_e/\sigma_x$

La correlación entre las puntuaciones empíricas y los errores vendrá dada por:

$$\begin{aligned}\rho_{xe} = \frac{\text{cov}\,(X, e)}{\sigma_x\sigma_e} = \frac{E(Xe) - E(X)E(e)}{\sigma_x\sigma_e} = \\ = \frac{E[(V + e)e] - E(V + e)E(e)}{\sigma_x\sigma_e} = \\ = \frac{E(Ve) + E(e^2) - E(V)E(e) - [E(e)]^2}{\sigma_x\sigma_e}\end{aligned}$$

Pero:

$$E(Ve) - E(V)E(e) = \sigma(V, e) = 0$$
$$E(e^2) - [E(e)]^2 = \sigma_e^2$$

Sustituyendo:

$$\rho_{xe} = \frac{\sigma_e^2}{\sigma_x \sigma_e}$$

Simplificando:

$$\rho_{xe} = \frac{\sigma_e}{\sigma_x}$$

1.9. $\mu_1 = \mu_2 = \cdots = \mu_k$

Para K formas paralelas de un test X, según el modelo:

$$X_1 = V + e_1;\ X_2 = V + e_2;$$
$$X_3 = V + e_3;\ \ldots\ ;\ X_k = V + e_k$$

Según 1.3:

$$\mu_{x_1} = \mu_v;\ \mu_{x_2} = \mu_v;\ \mu_{x_3} = \mu_v;\ \ldots\ ;\ \mu_{x_k} = \mu_v$$

Luego

$$\mu_{x_1} = \mu_{x_2} = \mu_{x_3} = \cdots = \mu_{x_k}$$

1.10. $\sigma_{x_1}^2 = \sigma_{x_2}^2 = \cdots = \sigma_{x_k}^2$

Para K formas paralelas de un test X, según 1.7:

$$\begin{aligned}
\sigma_{x_1}^2 &= \sigma_{v_1}^2 + \sigma_{e_1}^2 \\
\sigma_{x_2}^2 &= \sigma_{v_2}^2 + \sigma_{e_2}^2 \\
\sigma_{x_3}^2 &= \sigma_{v_3}^2 + \sigma_{e_3}^2 \\
&\vdots \quad\ \vdots \quad\ \vdots \\
\sigma_{x_k}^2 &= \sigma_{v_k}^2 + \sigma_{e_k}^2
\end{aligned}$$

Por definición de test paralelos:

$$\sigma_{e_1}^2 = \sigma_{e_2}^2 = \sigma_{e_3}^2 = \cdots = \sigma_{e_k}^2$$
$$\sigma_{v_1}^2 = \sigma_{v_2}^2 = \sigma_{v_3}^2 = \cdots = \sigma_{v_k}^2$$

Luego:

$$\sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma_{x_3}^2 = \cdots = \sigma_{x_k}^2$$

1.11. $\rho_{x_1x_2} = \rho_{x_1x_3} = \cdots = \rho_{x_jx_k}$

La correlación entre dos formas paralelas X_j, X_k vendrá dada por:

$$\rho_{x_jx_k} = \frac{\text{cov}\,(X_j, X_k)}{\sigma_{x_j}\sigma_{x_k}}$$

Ahora bien, según 1.6 la covarianza entre formas paralelas:

$$\text{cov}\,(X_j, X_k) = \sigma_v^2$$

y según 1.10:

$$\sigma_{x_j} = \sigma_{x_k}$$

Luego la correlación entre dos formas paralelas cualesquiera vendrá dada por:

$$\rho_{x_jx_k} = \frac{\sigma_v^2}{\sigma_x^2}$$

Pero según 1.10 las varianzas de las puntuaciones verdaderas (numerador) y de las empíricas (denominador) son iguales para todas las formas paralelas; luego $\rho_{x_jx_k}$ será constante para cualquier par j y k de formas paralelas, es decir:

$$\rho_{x_1x_2} = \rho_{x_1x_3} = \cdots = \rho_{x_jx_k}$$

2.1. $\rho_{xx'} = \sigma_v^2/\sigma_x^2$

La correlación entre dos formas paralelas X y X' vendrá dada por:

$$\rho_{xx'} = \frac{\text{cov}\,(X, X')}{\sigma_x \sigma_{x'}}$$

Según 1.6 (nota):

$$\text{cov}\,(X, X') = \sigma_v^2$$

Según 1.10

$$\sigma_x = \sigma_{x'}$$

luego

$$\sigma_x \sigma_{x'} = \sigma_x^2$$

Sustituyendo:

$$\rho_{xx'} = \frac{\sigma_v^2}{\sigma_x^2}$$

2.2. $\rho_{xx'} = 1 - [\sigma_e^2/\sigma_x^2]$

Despejando σ_v^2 de 1.7:

$$\sigma_v^2 = \sigma_x^2 - \sigma_e^2$$

Sustituyendo en 2.1

$$\rho_{xx'} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2}$$

dividiendo ambos términos entre σ_x^2,

$$\rho_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

2.3. $\rho_{xv} = \sqrt{\rho_{xx'}} = \sigma_v/\sigma_x$

La correlación entre X y V vendrá dada por:

$$\rho_{xv} = \frac{\text{cov}\,(X,V)}{\sigma_x \sigma_v}$$

Según 1.5:

$$\text{cov}\,(X,\, V) = \sigma_v^2$$

luego

$$\rho_{xv} = \frac{\sigma_v^2}{\sigma_x \sigma_v} = \frac{\sigma_v}{\sigma_x} = \sqrt{\frac{\sigma_v^2}{\sigma_x^2}} = \sqrt{\rho_{xx'}}$$

2.4. $\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}}$

Según 2.2:

$$\rho_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

operando:

$$\rho_{xx'} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2};\ \rho_{xx'} \cdot \sigma_x^2 = \sigma_x^2 - \sigma_e^2$$

$$\sigma_e^2 = \sigma_x^2 - \sigma_x^2 \rho_{xx'};\ \sigma_e^2 = \sigma_x^2 (1 - \rho_{xx'})$$

y extrayendo la raíz cuadrada:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}}$$

2.8. $Y' = \rho_{xy}(\sigma_y/\sigma_x)(X - \bar{X}) + \bar{Y}$

Sea:

$$y' = bx \text{ (puntuaciones diferenciales)}$$

Se trata de estimar el valor de b que minimice los errores de estimación $(y - y')$, y dado que $E(y - y') = 0$, que minimice una función de esos errores, los errores cuadráticos:

$$f(e) = E(y - y')^2$$

Desarrollando $f(e)$ y sustituyendo y' por su valor bx

$$\begin{aligned} f(e) = E(y - bx)^2 &= Ey^2 + b^2 Ex^2 - 2bExy = \\ &= \sigma_y^2 + b^2 \sigma_x^2 - 2b\,\text{cov}\,(x, y) = \\ &= \sigma_y^2 + b^2 \sigma_x^2 - 2b \rho_{xy} \sigma_y \sigma_x \end{aligned}$$

Derivando $f(e)$ respecto a b:

$$\frac{\delta f(e)}{\delta b} = 0 + 2b\sigma_x^2 - 2\rho_{xy}\sigma_y\sigma_x$$

Igualando a cero y despejando:

$$0 = 2b\sigma_x^2 - 2\rho_{xy}\sigma_y\sigma_x$$

$$b = \frac{2\rho_{xy}\sigma_y\sigma_x}{2\sigma_x^2} = \rho_{xy}\frac{\sigma_y}{\sigma_x}$$

Por tanto, en puntuaciones diferenciales:

$$y' = \rho_{xy}\frac{\sigma_y}{\sigma_x}x$$

Expresado en puntuaciones directas:

$$Y' - \bar{Y}' = \rho_{xy}\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

pero

$$\bar{Y}' = Y'$$

luego

$$Y' = \rho_{xy}\frac{\sigma_y}{\sigma_x}(X - \bar{X}) + \bar{Y}$$

2.10. $\sigma_{y\cdot x} = \sigma_y\sqrt{1 - \rho_{xy}^2}$

La varianza de los errores de estimación viene dada por:

$$\sigma^2{}_{y.x} = E(y - y')^2$$

Sustituyendo y' por su valor en diferenciales:

$$\sigma_{y\cdot x}^2 = E\left[y - \rho_{xy}\left(\frac{\sigma_y}{\sigma_x}\right)x\right]^2$$

$$\sigma_{y\cdot x}^2 = Ey^2 + \rho_{xy}^2\left(\frac{\sigma_y^2}{\sigma_x^2}\right)Ex^2 - 2\rho_{xy}\left(\frac{\sigma_y}{\sigma_x}\right)Exy$$

Ahora bien:

$$Ex^2 = \sigma_x^2;\ Exy = \text{cov}\,(x, y) = \rho_{xy}\sigma_y\sigma_x$$

luego

$$\sigma_{y\cdot x}^2 = \sigma_y^2 + \rho_{xy}^2\left(\frac{\sigma_y^2}{\sigma_x^2}\right)\sigma_x^2 - 2\rho_{xy}\left(\frac{\sigma_y}{\sigma_x}\right)\rho_{xy}\sigma_y\sigma_x$$

Simplificando:

$$\sigma_{y\cdot x}^2 = \sigma_y^2 + \rho_{xy}^2\sigma_y^2 - 2\rho_{xy}^2\sigma_y^2 =$$
$$= \sigma_y^2 - \rho_{xy}^2\sigma_y^2 = \sigma_y^2(1 - \rho_{xy}^2)$$

Extrayendo la raíz cuadrada:

$$\sigma_{y\cdot x} = \sigma_y\sqrt{1 - \rho_{xy}^2}$$

2.11. $\sigma_{v\cdot x} = \sigma_x\sqrt{1 - \rho_{xx'}}\sqrt{\rho_{xx'}}$

La fórmula general viene dada en 2.10.

$$\sigma_{y\cdot x} = \sigma_y\sqrt{1 - \rho_{xy}^2}$$

En el modelo Y pasa a ser V:

$$\sigma_{v\cdot x} = \sigma_v\sqrt{1 - \rho_{vx}^2}$$

Ahora bien, según 2.3:

$$\rho_{xv} = \frac{\sigma_v}{\sigma_x}$$

luego

$$\sigma_v = \rho_{xv}\sigma_x$$

Sustituyendo:

$$\rho_{v.x} = \rho_{xv}\sigma_x\sqrt{1 - \rho_{xv}^2}$$

Pero

$$\rho_{xv} = \sqrt{\rho_{xx'}} \quad \text{o} \quad \rho_{xv}^2 = \rho_{xx'}$$

Sustituyendo:

$$\sigma_{v.x} = \sigma_x\sqrt{1-\rho_{xx'}}\sqrt{\rho_{xx'}}$$

2.13. $$\rho_{dd'} = \frac{\sigma_x^2\rho_{xx'} + \sigma_z^2\rho_{zz'} - 2\sigma_x\sigma_z\rho_{xz}}{\sigma_x^2 + \sigma_z^2 - 2\sigma_x\sigma_z\rho_{xz}}$$

Sea d la diferencia entre las puntuaciones X y Z:

$$(d = X - Z)$$

El coeficiente de fiabilidad de las puntuaciones d vendrá dado por:

$$\rho_{dd'} = \frac{\sigma_v^2}{\sigma_x^2} = \frac{\sigma^2_{(V_x - V_z)}}{\sigma^2_{(X-Z)}} =$$

$$= \frac{\sigma^2_{v_x} + \sigma^2_{v_z} - 2\,\mathrm{cov}\,(V_x, V_z)}{\sigma_x^2 + \sigma_z^2 - 2\,\mathrm{cov}\,(X, Z)}$$

Ahora bien:

$$\rho_{xx'} = \frac{\sigma_v^2}{\sigma_x^2}$$

luego

$$\sigma^2_{vx} = \rho_{xx'}\sigma_x^2$$

además, según 1.6:

$$\mathrm{cov}\,(V_x, V_z) = \mathrm{cov}\,(X, Z) = \sigma_x\sigma_z\rho_{xz}$$

sustituyendo:

$$\rho_{dd'} = \frac{\sigma_x^2\rho_{xx'} + \sigma_z^2\rho_{zz'} - 2\sigma_x\sigma_z\rho_{xz}}{\sigma_x^2 + \sigma_z^2 - 2\sigma_x\sigma_z\rho_{xz}}$$

2.14. $$\rho_{dd'} = \frac{\rho_{xx'} + \rho_{zz'} - 2\rho_{xz}}{2(1-\rho_{xz})}$$

Si los test están en la misma escala, entonces en 2.13:

$$\sigma_x^2 = \sigma_z^2 \quad \text{y} \quad \sigma_x\sigma_z = \sigma_x^2$$

luego sacando factor común σ_x^2:

$$\rho_{dd'} = \frac{\sigma_x^2(\rho_{xx'} + \rho_{zz'} - 2\rho_{xz})}{\sigma_x^2(2 - 2\rho_{xz})}$$

simplificando:

$$\rho_{dd'} = \frac{\rho_{xx'} + \rho_{zz'} - 2\rho_{xz}}{2(1-\rho_{xz})}$$

2.16. $$\sigma_{e_s} = \sigma_x\sqrt{1-\rho_{xx'}}\sqrt{2}$$

$e = x_1 - x_2$ (en puntuaciones diferenciales)

$$\sigma^2_{e_s} = E(x_1 - x_2)^2$$
$$\sigma^2_{e_s} = E(x_1^2 + x_2^2 - 2x_1x_2)$$
$$\sigma^2_{e_s} = Ex_1^2 + Ex_2^2 - 2Ex_1x_2$$
$$\sigma^2_{e_s} = \sigma^2_{x_1} + \sigma^2_{x_2} - 2\,\mathrm{cov}\,(x_1, x_2)$$
$$\sigma^2_{e_s} = \sigma^2_{x_1} + \sigma^2_{x_2} - 2\rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2}$$

Teniendo en cuenta que x_1 y x_2 son formas paralelas:

$$\sigma^2_{x_1} = \sigma^2_{x_2}$$

luego

$$\sigma^2_{e_s} = 2\sigma_x^2 - 2\rho_{xx'}\sigma_x^2$$

y sacando factor común $2\sigma_x^2$:

$$\sigma^2_{e_s} = 2\sigma_x^2(1 - \rho_{xx'})$$

Extrayendo la raíz cuadrada:

$$\sigma_{e_s} = \sigma_x\sqrt{1-\rho_{xx'}}\sqrt{2}$$

2.17. $$\sigma_{e_p} = \sigma_x\sqrt{1-\rho_{xx'}}\sqrt{1+\rho_{xx'}}$$

$e = x_1 - x'_1$ (en puntuaciones diferenciales). Los pronósticos de x_1 a partir de x_2 vendrán dados por:

$$x_1' = \rho_{x_1x_2}\left(\frac{\sigma_{x_1}}{\sigma_{x_2}}\right)x_2$$

$$\sigma^2_{e_p} = E(x_1 - x_1')^2 = E\left[x_1 - \rho_{x_1x_2}\left(\frac{\sigma_{x_1}}{\sigma_{x_2}}\right)x_2\right]^2$$

$$\sigma^2_{e_p} = Ex_1^2 + \rho^2_{x_1x_2}\left(\frac{\sigma^2_{x_1}}{\sigma^2_{x_2}}\right)Ex_2^2 - 2\rho_{x_1x_2}\left(\frac{\sigma_{x_1}}{\sigma_{x_2}}\right)Ex_1x_2$$

Teniendo en cuenta que x_1 y x_2 son formas paralelas:

$$\sigma_{x_1} = \sigma_{x_2} = \sigma_x \quad ; \quad \rho_{x_1x_2} = \rho_{xx'}$$

luego:

$$\sigma^2_{e_p} = \sigma^2_{x_1} + \rho^2_{xx'}\left(\frac{\sigma^2_{x_1}}{\sigma^2_{x_2}}\right)\sigma^2_{x_2} - 2\rho_{xx'}\left(\frac{\sigma_{x_1}}{\sigma_{x_2}}\right)\rho_{xx'}\sigma_{x_1}\sigma_{x_2}$$

Simplificando

$$\sigma^2_{e_p} = \sigma^2_x + \rho^2_{xx'}\sigma^2_x - 2\rho^2_{xx'}\sigma^2_x = \sigma^2_x - \rho^2_{xx'}\sigma^2_x$$
$$\sigma^2_{e_p} = \sigma^2_x(1 - \rho^2_{xx'}) = \sigma^2_x(1 + \rho_{xx'})(1 - \rho_{xx'})$$

Extrayendo la raíz cuadrada:

$$\sigma_{e_p} = \sigma_x\sqrt{1 - \rho_{xx'}}\sqrt{1 + \rho_{xx'}}$$

2.20. $$\rho_{xx'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}}$$

Hay varias posibles formas de derivar 2.20. Veamos una muy sencilla. La fiabilidad viene dada por $\rho_{xx'} = \sigma^2_v/\sigma^2_x$; por tanto, si se desea saber el efecto que tiene sobre $\rho_{xx'}$ alargar n veces paralelas el test, veamos lo que ocurre a los factores de los que depende $\rho_{xx'}$, esto es, σ^2_v y σ^2_x. Si se alarga el test n veces:

$$V = v_1 + v_2 + \cdots + v_n$$
$$X = x_1 + x_2 + \cdots + x_n$$

Calculemos las varianzas de V y X en función de sus componentes:

$$\sigma^2_V = \sigma^2(v_1 + v_2 + \cdots + v_n) =$$
$$= \sum \sigma^2_{vj} + \sum\sum \text{cov}\,(v_j, v_k)$$

Ahora bien, para tests paralelos y puntuaciones verdaderas:

$$\sigma_{v_j} = \sigma_{v_k} = \sigma_v \quad \text{y} \quad \text{cov}\,(v_j, v_k) = \sigma^2_v$$

luego

$$\sigma^2_V = n\sigma^2_v + n(n-1)\sigma^2_v = \sigma^2_v[n + n(n-1)] =$$
$$= \sigma^2_v(n + n^2 - n)$$

simplificando:

$$\sigma^2_V = n^2\sigma^2_v$$

Varianza de X:

$$\sigma^2_X = \sigma^2(x_1 + x_2 + \cdots + x_n) =$$
$$= \sum \sigma^2_{xj} + \sum\sum \text{cov}\,(x_j, x_k)$$

donde

$$\sigma_{xj} = \sigma_{xk} = \sigma_x \quad ;$$
$$\text{cov}\,(x_j, x_k) = \rho_{x_jx_k}\sigma_{x_j}\sigma_{x_k} = \rho_{xx'}\sigma^2_x$$

luego:

$$\sigma^2_X = n\sigma^2_x + n(n-1)\rho_{xx'}\sigma^2_x$$
$$\sigma^2_X = n\sigma^2_x[1 + (n-1)\rho_{xx'}]$$

Por tanto:

$$\rho_{XX'} = \frac{\sigma^2_V}{\sigma^2_X} = \frac{n^2\sigma^2_v}{n\sigma^2_x[1 + (n-1)\rho_{xx'}]}$$

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}}$$

2.23. $$\alpha = \frac{n}{n-1}\left[1 - \left(\frac{\sum \sigma_j^2}{\sigma_x^2}\right)\right]$$

Según 2.20:

$$\rho_{XX'} = \frac{\sigma_V^2}{\sigma_X^2} = \frac{n^2\sigma_v^2}{n\sigma_x^2 + n(n-1)\rho_{xx'}\sigma_x^2}$$

Si los componentes son paralelos:

$$\sigma_v^2 = \rho_{x_j x_k}\sigma_{x_j}^2 = \rho_{xx'}\sigma_j^2$$
$$\sigma_X^2 = \sum \sigma_{x_j}^2 + n(n-1)\rho_{x_j x_k}\sigma_{x_j}^2 =$$
$$= \sum \sigma_j^2 + n(n-1)\rho_{xx'}\sigma_j^2$$

Sustituyendo:

$$\rho_{XX'} = \frac{n^2\rho_{xx'}\sigma_j^2}{\sigma_X^2}$$

Multiplicando numerador y denominador por $(n-1)$:

$$\rho_{XX'} = \frac{nn(n-1)\rho_{xx'}\sigma_j^2}{(n-1)\sigma_X^2}$$

Ahora bien:

$$n(n-1)\rho_{xx'}\sigma_j^2 = \sigma_X^2 - \sum \sigma_j^2$$

luego

$$\rho_{XX'} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_X^2 - \sum \sigma_j^2}{\sigma_X^2}\right)$$
$$\rho_{XX'} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum \sigma_j^2}{\sigma_X^2}\right) = \alpha$$

Por tanto, cuando los componentes son paralelos $\alpha = \rho_{XX'}$.

2.24. $$\alpha = \left(\frac{n}{n-1}\right)\left(\frac{\sum\sum \text{cov}\,(j,k)}{\sigma_X^2}\right)$$

La varianza de X en función de sus componentes:

$$\sigma_X^2 = \sum \sigma_j^2 + \sum\sum \text{cov}\,(j,k)$$

de donde

$$\sum\sum \text{cov}\,(j,k) = \sigma_X^2 - \sum \sigma_j^2$$

Ahora bien, según 2.23 (penúltimo paso):

$$\alpha = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_X^2 - \sum \sigma_j^2}{\sigma_X^2}\right)$$

Sustituyendo:

$$\sigma_X^2 - \sum \sigma_j^2 = \sum\sum \text{cov}\,(j,k)$$

nos queda

$$\alpha = \left(\frac{n}{n-1}\right)\left(\frac{\sum\sum \text{cov}\,(j,k)}{\sigma_X^2}\right)$$

2.25. $$\bar{\alpha} \underset{n\to\infty}{=} \hat{\alpha}$$

$$\bar{\alpha} = \frac{(N-3)\hat{\alpha}}{N-1} + \frac{2}{N-1} = \frac{N\hat{\alpha} - 3\hat{\alpha}}{N-1} + \frac{2}{N-1} =$$
$$= \frac{N\hat{\alpha}}{N-1} - \frac{3\hat{\alpha}}{N-1} + \frac{2}{N-1} =$$
$$= \frac{\hat{\alpha}}{\frac{(N-1)}{N}} - \frac{3\hat{\alpha}}{N-1} + \frac{2}{N-1} =$$
$$= \frac{\hat{\alpha}}{\frac{N}{N} - \frac{1}{N}} - \frac{3\hat{\alpha}}{N-1} = \frac{\hat{\alpha}}{1 - \frac{1}{N}} - \frac{3\hat{\alpha}}{N-1} + \frac{2}{N-1}$$

Cuando $N \to \infty$

$$\frac{1}{N} = 0 \quad ; \quad \frac{3\alpha}{N-1} = 0 \quad ; \quad \frac{2}{N-1} = 0$$

luego

$$\bar{\alpha} = \hat{\alpha}$$

2.26. $\alpha \leqslant \rho_{XX'}$

Caso de dos componentes

Sean X_1 y X_2 las puntuaciones empíricas y V_1 y V_2 las puntuaciones verdaderas:

1.
$$(\sigma_{v_1} - \sigma_{v_2})^2 \geqslant 0$$
desarrollando:

2.
$$\sigma_{v_1}^2 + \sigma_{v_2}^2 - 2\sigma_{v_1}\sigma_{v_2} \geqslant 0$$
luego
$$\sigma_{v_1}^2 + \sigma_{v_2}^2 \geqslant 2\sigma_{v_1}\sigma_{v_2}$$
ahora bien,

3.
$$\rho_{v_1v_2} = \sigma_{v_1v_2}/(\sigma_{v_1}\sigma_{v_2})$$
puesto que $\rho_{v_1v_2} \leqslant 1$:
$$\sigma_{v_1v_2}/(\sigma_{v_1}\sigma_{v_2}) \leqslant 1$$
luego:

4.
$$\sigma_{v_1v_2} \leqslant \sigma_{v_1}\sigma_{v_2}$$

5. Según 2 y 4:
$$\sigma_{v_1}^2 + \sigma_{v_2}^2 \geqslant 2\sigma_{v_1v_2}$$
sumando $2\sigma_{v_1v_2}$ a 5:

6.
$$\sigma_{v_1}^2 + \sigma_{v_2}^2 + 2\sigma_{v_1v_2} \geqslant 4\sigma_{v_1v_2}$$

7. Ahora bien:
$$\sigma_{v_1}^2 + \sigma_{v_2}^2 + 2\sigma_{v_1v_2} = \sigma_V^2$$
luego
$$\sigma_V^2 \geqslant 4\sigma_{v_1v_2}$$

8. Dividiendo ambos miembros de 7 entre σ_X^2
$$\sigma_V^2/\sigma_X^2 \geqslant 4\sigma_{v_1v_2}/\sigma_X^2$$

9. Ahora bien,
$$\sigma_V^2/\sigma_X^2 = \rho_{XX'} \quad \text{y} \quad \sigma_{v_1v_2} = \sigma_{xx'}$$
según [1.6], luego:
$$\rho_{XX'} \geqslant 4\sigma_{x_1x_2}/\sigma_X^2 = 2(2\sigma_{x_1x_2}/\sigma_X^2)$$

10. Teniendo en cuenta que
$$\sigma_X^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 + 2\sigma_{x_1x_2}$$
$$2\sigma_{x_1x_2} = \sigma_X^2 - (\sigma_{x_1}^2 + \sigma_{x_2}^2)$$
Sustituyendo en 9:
$$\rho_{XX'} \geqslant 2\left[\frac{\sigma_X^2 - (\sigma_{x_1}^2 + \sigma_{x_2}^2)}{\sigma_X^2}\right]$$
operando:
$$\rho_{XX'} \geqslant \frac{2}{2-1}\left[1 - \left(\frac{\sigma_{x_1}^2 + \sigma_{x_2}^2}{\sigma_X^2}\right)\right]$$

11. Pero
$$\frac{2}{2-1}\left[1 - \left(\frac{\sigma_{x_1}^2 + \sigma_{x_2}^2}{\sigma_X^2}\right)\right] = \alpha$$
luego:
$$\rho_{XX'} \geqslant \alpha$$

Caso general de n componentes

Según la desigualdad de Cauchy-Schwartz:
$$\sum\sum (\sigma_{v_j}^2 + \sigma_{v_k}^2) \geqslant 2\sum\sum \sigma_{v_jv_k}$$
Ahora bien,
$$\sum\sum (\sigma_{v_j}^2 + \sigma_{v_k}^2) = 2(n-1)\sum \sigma_{v_j}^2$$
(véase *nota* al final)
luego:
$$2(n-1)\sum \sigma_{v_j}^2 \geqslant 2\sum\sum \sigma_{v_jv_k}$$
donde:
$$\sum \sigma_{v_j}^2 \geqslant \frac{2\sum\sum \sigma_{v_jv_k}}{2(n-1)}$$

Sumando a ambos miembros $\sum\sum \sigma_{v_j v_k}$

$$\sum \sigma_{v_j}^2 + \sum\sum \sigma_{v_j v_k} \geqslant \frac{\sum\sum \sigma_{v_j v_k}}{n-1} + \sum\sum \sigma_{v_j v_k}$$

El primer miembro es σ_V^2; luego:

$$\sigma_V^2 \geqslant \frac{\sum\sum \sigma_{v_j v_k}}{n-1} + \frac{\sum\sum \sigma_{v_j}\sigma_{v_k}(n-1)}{n-1}$$

$$\sigma_V^2 \geqslant \frac{n\sum\sum \sigma_{v_j}\sigma_{v_k}}{n-1}$$

$$\sigma_V^2 \geqslant \left(\frac{n}{n-1}\right)\sum\sum \sigma_{v_j}\sigma_{v_k}$$

Dividiendo entre σ_X^2 y sustituyendo $\sum\sum \sigma_{v_j v_k}$ por su equivalente $(\sigma_X^2 - \sum \sigma_j^2)$:

$$\frac{\sigma_V^2}{\sigma_X^2} \geqslant \left(\frac{n}{n-1}\right)\left(\frac{\sigma_X^2 - \sum \sigma_j^2}{\sigma_X^2}\right)$$

$$\rho_{XX'} \geqslant \left(\frac{n}{n-1}\right)\left(\frac{\sigma_X^2 - \sum \sigma_j^2}{\sigma_X^2}\right)$$

pero el segundo miembro de la desigualdad es α; luego:

$$\rho_{XX'} \geqslant \alpha$$

Nota. Aunque es fácil de ver que según las propiedades del sumatorio

$$\sum\sum (\sigma_{v_j}^2 + \sigma_{v_k}^2) = 2(n-1)\sum \sigma_{v_j}^2$$

Lord y Novick (1968, p. 89) ofrecen una sencilla demostración:

a) $\sum\sum(\sigma_{v_j}^2+\sigma_{v_k}^2) = n\sum\sigma_{v_j}^2 + n\sum\sigma_{v_k}^2 = 2n\sum\sigma_{v_j}^2$

b) $\sum\sum(\sigma_{v_j}^2+\sigma_{v_k}^2) = \sum\sum(\sigma_{v_j}^2+\sigma_{v_k}^2) + \sum\sum(\sigma_{v_j}^2+\sigma_{v_k}^2) = 2\sum\sigma_{v_j}^2 + \sum\sum(\sigma_{v_j}^2+\sigma_{v_k}^2)$

Igualando *a*) y *b*) y despejando:

$$2n\sum\sigma_{v_j}^2 = 2\sum\sigma_{v_j}^2 + \sum\sum(\sigma_{v_j}^2 + \sigma_{v_k}^2)$$

$$2(n-1)\sum\sigma_{v_j}^2 = \sum\sum(\sigma_{v_j}^2 + \sigma_{v_k}^2)$$

2.34. Utilizando el análisis de varianza α viene dado por:

$$\alpha = \frac{MC_{\text{personas}} - MC_{\text{residual}}}{MC_{\text{personas}}} = 1 - \frac{MC_r}{MC_p}$$

luego:

$$1 - \hat{\alpha} = \frac{MC_r}{MC_p}$$

Análogamente, en la población:

$$1 - \alpha = \frac{E(MC_r)}{E(MC_p)}$$

Dividiendo miembro a miembro:

$$\frac{1-\alpha}{1-\hat{\alpha}} = \frac{\dfrac{E(MC_r)}{E(MC_p)}}{\dfrac{MC_r}{MC_p}}$$

reordenando:

$$\frac{1-\alpha}{1-\hat{\alpha}} = \frac{\dfrac{MC_p}{E(MC_p)}}{\dfrac{MC_r}{E(MC_r)}}$$

Ahora bien, en el modelo ANOVA numerador y denominador son independientes y se distribuyen según χ^2 con $(N-1)$ y $(n-1)(N-1)$ grados de libertad, respectivamente; luego su cociente se distribuye según $F_{(N-1),(n-1)(N-1)}$. (Se asumen los supuestos del modelo ANOVA tipo II con una observación por casilla.)

2.35. W no se distribuye estrictamente según $F_{(N_1-1),(N_2-1)}$ sino según $(F_1)(F_2)$ con grados de libertad, respectivamente, de:

$$[(N_1-1),(N_1-1)(n_1-1)] \quad \text{y}$$
$$[(N_2-1),(N_2-1)(n_2-1)]$$

donde:

N: Número de sujetos.
n: Número de ítems.

Feld (1969) demuestra que en muestras grandes $(F_1)(F_2)$ se aproxima a $F_{(N_1-1),(N_2-1)}$, entendiendo por grandes $N > 100$.

2.39. *Media.*

Es inmediato que si una variable X es combinación lineal de n variables $x_1, x_2, ..., x_n$ con pesos $a_1, a_2, ..., a_n$, su media viene dada por:

$$\bar{X} = \mathbf{a}_i\mathbf{u}$$

donde $\mathbf{a}_i$ es el vector de pesos y $\mathbf{u}$ el vector de medias de los componentes. Veámoslo: Sea

$$X = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

O, en forma matricial:

$$\bar{X} = \mathbf{a}'\mathbf{u}$$

la media de X vendrá dada por:

$$E(X) = E(\mathbf{a}'\mathbf{x}) = \mathbf{a}'E(\mathbf{x}) = \mathbf{a}'\mathbf{u}$$

donde $\mathbf{u}$ es el vector de medias

$$\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_n$$

Varianza

$$\begin{aligned}\text{var}\,(X) = \text{var}\,(\mathbf{a}'\mathbf{x}) &= E(\mathbf{a}'\mathbf{x} - \mathbf{a}'\mathbf{u})^2 = \\ &= E(\mathbf{a}'\mathbf{x} - \mathbf{a}'\mathbf{u})(\mathbf{a}'\mathbf{x} - \mathbf{a}'\mathbf{u}) = \\ &= E[\mathbf{a}'(\mathbf{x} - \mathbf{u})(\mathbf{x} - \mathbf{u})'\mathbf{a}] = \\ &= \mathbf{a}'[E(\mathbf{x} - \mathbf{u})(\mathbf{x} - \mathbf{u})']\mathbf{a}\end{aligned}$$

Pero

$$E(\mathbf{x} - \mathbf{u})(\mathbf{x} - \mathbf{u})'$$

es la matriz de varianzas-covarianzas de $\mathbf{x}$: Σ_{xx}; luego

$$\text{var}\,(X) = \mathbf{a}'\Sigma_{xx}\,\mathbf{a}$$

2.40. Las esperanzas matemáticas de las medias cuadráticas vienen dadas por:

Personas (p)	$\sigma_e^2 + n_m\sigma_{pi}^2 + n_i\sigma_{pm}^2 + n_in_m\sigma_p^2$
Evaluadores (i)	$\sigma_e^2 + n_pv_{im}^2 + n_mv_{pi}^2 + n_pn_mv_i^2$
Modalidad (m)	$\sigma_e^2 + n_pv_{im}^2 + n_iv_{pm}^2 + n_pn_mv_m^2$
$p \times i$	$\sigma_e^2 + n_mv_{pi}^2$
$p \times m$	$\sigma_e^2 + n_i\sigma_{pm}^2$
$i \times m$	$\sigma_e^2 + n_pv_{im}^2$
$p \times i \times m + e$	σ_e^2

Dado que ya se habían calculado las medias cuadráticas de las distintas fuentes de variación (véase la tabla 2.7), para conocer el valor de las varianzas se despejan de la ecuación correspondiente:

$$\sigma_e^2 = 1{,}050$$

$$v_{im}^2 = \frac{2{,}40 - 1{,}05}{10} = 0{,}135$$

$$\sigma_{pm}^2 = \frac{2{,}650 - 1{,}050}{2} = 1{,}210$$

$$v_m^2 = \frac{9{,}82 - 1{,}050 - 10(0{,}135) - 3(0{,}533)}{10 \times 3} = 0{,}194$$

$$v_i^2 = \frac{5{,}35 - 1{,}050 - 10(0{,}135) - 2(1{,}210)}{10 \times 2} = 0{,}026$$

$$\sigma_p^2 = \frac{7{,}84 - 1{,}050 - 2(1{,}210) - 3(0{,}533)}{3 \times 2} = 0{,}462$$

3.20. $\sigma_y^2 = \sigma_{y'}^2 + \sigma_{y.x}^2$

$$\begin{aligned}\sigma_y^2 = E(Y - \bar{Y})^2 &= E(Y - \bar{Y} + Y' - Y') = \\ &= E[(Y' - \bar{Y}) + (Y - Y')]^2 = \\ &= E[(Y' - \bar{Y})^2 + (Y - Y')^2 + 2(Y' - \bar{Y})(Y - Y')] = \\ &= E(Y' - \bar{Y})^2 + E(Y - Y')^2 + 2E(Y' - \bar{Y})(Y - Y')\end{aligned}$$

Ahora bien:

$$E(Y' - \bar{Y})^2 = E(Y' - \bar{Y}')^2$$

puesto que

$$\bar{Y} = \bar{Y}'$$

Además:

$$E(Y' - \bar{Y})(Y - Y') = \text{cov}\,[(Y' - \bar{Y}),(Y - Y')] = 0$$

Luego:

$$\sigma_y^2 = E(Y' - \bar{Y}')^2 + E(Y - Y')^2$$
$$\sigma_y^2 = \sigma_{y'}^2 + \sigma_{y.x}^2$$

3.21. $\rho_{xy}^2 = \dfrac{\sigma_{y'}^2}{\sigma_y^2}$

$$y' = \frac{\rho_{xy}\sigma_y}{\sigma_x} x$$

luego:

$$\sigma_{y'}^2 = \frac{\rho_{xy}^2\sigma_y^2}{\sigma_x^2}\sigma_x^2 = \rho_{xy}^2\sigma_y^2$$

sustituyendo:

$$\rho_{xy}^2 = \frac{\sigma_{y'}^2}{\sigma_y^2} = \frac{\rho_{xy}^2\sigma_y^2}{\sigma_y^2} = \rho_{xy}^2$$

3.22. $\rho_{xy}^2 = 1 - \dfrac{\sigma_{y.x}^2}{\sigma_y^2}$

Dividiendo 3.20 entre σ_y^2:

$$\frac{\sigma_y^2}{\sigma_y^2} = \frac{\sigma_{y'}^2}{\sigma_y^2} + \frac{\sigma_{y.x}^2}{\sigma_y^2}$$

Teniendo en cuenta 3.21 y simplificando:

$$1 = \rho_{xy}^2 + \frac{\sigma_{y.x}^2}{\sigma_y^2}$$

$$\rho_{xy}^2 = 1 - \frac{\sigma_{y.x}^2}{\sigma_y^2}$$

3.24. $\mathbf{b} = (XX')^{-1}X'\mathbf{Y}$

Se trata de minimizar $\mathbf{e}'\mathbf{e}$.
Según el modelo:

$$\mathbf{Y} = X\mathbf{b} + e$$

de donde:

$$\mathbf{e} = \mathbf{Y} - X\mathbf{b}$$

por tanto:

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= (\mathbf{Y} - X\mathbf{b})'(\mathbf{Y} - X\mathbf{b}) = \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'X\mathbf{b} - \mathbf{b}'X'\mathbf{Y} + \mathbf{b}'X'X\mathbf{b} = \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'X\mathbf{b} - (\mathbf{Y}'X\mathbf{b})' + \mathbf{b}'X'X\mathbf{b} = \\ &= \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'X\mathbf{b} + \mathbf{b}'X'X\mathbf{b}\end{aligned}$$

Ya que $\mathbf{Y}'X\mathbf{b} = (\mathbf{Y}'X\mathbf{b})' = \mathbf{Y}'X\mathbf{b}$ por ser un escalar, y por tanto igual a su traspuesta, derivando $\mathbf{e}'\mathbf{e}$ respecto a $\mathbf{b}$:

$$\frac{\delta\mathbf{e}'\mathbf{e}}{\delta\mathbf{b}} = 0 - 2X'\mathbf{Y} + 2(X'X)\mathbf{b}$$

igualando a cero:

$$0 = X'\mathbf{Y} - (X'\mathbf{X})\mathbf{b}$$

despejando **b:**

$$\mathbf{b} = (X'X)^{-1}X'\mathbf{Y}$$

Propiedades de **b**

1. $E(\mathbf{b}) = \mathbf{B}$

$$E(\mathbf{b}) = E[(X'X)^{-1}X'\mathbf{Y}] = E[(X'X)^{-1}X'(X\mathbf{B} + \mathbf{e})]$$

Ahora bien, en el modelo:

$$\mathbf{Y} = X\mathbf{B} + \mathbf{e}$$

luego:

$$E(\mathbf{b}) = E[(X'X)^{-1}X'\mathbf{Y}] = E[(X'X)^{-1}X'(X\mathbf{B} + \mathbf{e})]$$

Operando:

$$E(\mathbf{b}) = (X'X)^{-1}X'XE(\mathbf{B}) + (X'X)^{-1}X'E(\mathbf{e})$$

Ahora bien:

$$E(\mathbf{e}) = 0 \quad ; \quad (X'X)^{-1}X'X = I$$

luego:

$$E(\mathbf{b}) = IE(\mathbf{B}) + 0 = E(\mathbf{B}) = B$$

2. $\text{var}\,(\mathbf{b}) = \sigma_e^2(X'X)^{-1}$

$$\begin{aligned}\text{var}\,(\mathbf{b}) &= E\{[\mathbf{b} - \mu(\mathbf{b})][\mathbf{b} - \mu(\mathbf{b})]'\} = \\ &= E[(\mathbf{b} - \mathbf{B})(\mathbf{b} - \mathbf{B})']\end{aligned}$$

pero:

$$\begin{aligned}(\mathbf{b} - \mathbf{B}) &= (X'X)^{-1}X'\mathbf{Y} - \mathbf{B} = \\ &= (X'X)^{-1}X'(X\mathbf{B} + e) - \mathbf{B} = \\ &= (X'X)^{-1}X'X\mathbf{B} + (X'X)^{-1}X'\mathbf{e} - \mathbf{B} = \\ &= \mathbf{B} + (X'X)^{-1}X'\mathbf{e} - \mathbf{B} = (X'X)^{-1}X'\mathbf{e}\end{aligned}$$

luego:

$$\begin{aligned}\text{var}\,(\mathbf{b}) &= E\{[(X'X)^{-1}X'\mathbf{e}][(X'X)^{-1}X'\mathbf{e}]'\} = \\ &= \mathrm{E}\{[(X'X)^{-1}X'\mathbf{e}][\mathbf{e}'X(X'X)^{-1}]\} = \\ &= (X'X)^{-1}X'E(\mathbf{e}\mathbf{e}')X(X'X)^{-1}\end{aligned}$$

pero $E(\mathbf{e}\mathbf{e}') = \sigma_e^2 I$; luego:

$$\text{var}\,(\mathbf{b}) = \sigma_e^2(X'X)^{-1}X'X(X'X)^{-1}$$

donde:

$$(X'X)^{-1}X'X = I$$

luego:

$$\text{var}\,(\mathbf{b}) = \sigma_e^2(X'X)^{-1}$$

3.33. *Estimador insesgado de* $\sigma^2_{y.\,123\ldots\,k}$

Es fácil demostrar que:

$$E(\mathbf{e}'\mathbf{e}) = \sigma^2_{y.\,123\ldots\,k}(N - K - 1)$$

o lo que es lo mismo:

$$E\left[\frac{\mathbf{e}'\mathbf{e}}{N - K - 1}\right] = \sigma^2_{y.123\ldots k}$$

donde:

$$\mathbf{e}'\mathbf{e} = (\mathbf{Y} - X\mathbf{b})'(\mathbf{Y} - X\mathbf{b}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'X'\mathbf{Y}$$

sustituyendo:

$$E\left[\frac{\mathbf{Y}'\mathbf{Y} - \mathbf{b}'X'\mathbf{Y}}{N - K - 1}\right] = \sigma^2_{y.123\ldots k}$$

es decir:

$$\frac{\mathbf{Y}'\mathbf{Y} - \mathbf{b}'X'\mathbf{Y}}{N - K - 1}$$

es un estimador insesgado de

$$\sigma^2_{y.\,123\ldots\,k}$$

3.37.

$$r_{xy.z} = \frac{r_{xy} - r_{zy}r_{zx}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{zy}^2}}$$

Según la definición de la correlación parcial:

$$r_{xy.z} = r_{(x-x')(y-y')} = \frac{\sum (x - x')(y - y')}{NS_{(x-x')}S_{(y-y')}} =$$

$$= \frac{\sum xy - \sum xy' - \sum x'y - \sum x'y'}{NS_x\sqrt{1 - r_{zx}^2}S_y\sqrt{1 - r_{zy}^2}}$$

Ahora bien:

$$y' = \frac{r_{zy}S_y}{S_z}$$

$$x' = \frac{r_{zx}S_x}{S_z}$$

Sustituyendo y operando:

$$r_{xy.z} = \frac{\sum xy - \left(\frac{r_{zy}S_y}{S_z}\right)\sum zx - \left(\frac{r_{zx}S_x}{S_z}\right)\sum zy + \left(\frac{r_{zx}S_x}{S_z}\right)\left(\frac{r_{zy}S_y}{S_z}\right)\sum z^2}{NS_x\sqrt{1 - r_{zx}^2}S_y\sqrt{1 - r_{zy}^2}}$$

Teniendo en cuenta los valores de r_{xy}, r_{zx} y r_{zy} y simplificando:

$$r_{xy.z} = \frac{r_{xy} - r_{zy}r_{zx} - r_{zx}r_{zy} + r_{zx}r_{zy}}{\sqrt{1 - r_{zx}^2}\sqrt{1 - r_{zy}^2}} =$$

$$= \frac{r_{xy} - r_{zy}r_{zx}}{\sqrt{1 - r_{zx}^2}\sqrt{1 - r_{zy}^2}}$$

3.39.

$$r_{(x-x')y} = \frac{r_{xy} - r_{zy}}{\sqrt{1 - r_{zx}^2}}$$

$$r_{(x-x')y} = \sum \frac{(x - x')y}{NS_{(x-x')}S_y} = \frac{\sum xy - \sum x'y}{NS_x\sqrt{1 - r_{zx}^2}S_y}$$

Teniendo en cuenta que

$$r_{xy} = \frac{\sum xy}{NS_xS_y}$$

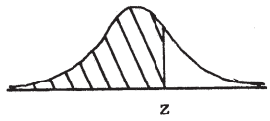
$$r_{zy} = \frac{\sum zy}{NS_zS_y}$$

y simplificando:

$$r_{(x-x')y} = \frac{r_{xy} - r_{zx}r_{zy}}{\sqrt{1 - r_{zx}^2}}$$

Tablas estadísticas

TABLA A. Distribución normal. $P(Z \leq z)$.



z	0	1	2	3	4	5	6	7	8	9
-3'5	'0002	'0002	'0002	'0002	'0002	'0002	'0002	'0002	'0002	'0002
-3'4	'0003	'0003	'0003	'0003	'0003	'0003	'0003	'0003	'0002	'0002
-3'3	'0005	'0005	'0004	'0004	'0004	'0004	'0004	'0004	'0004	'0003
-3'2	'0007	'0007	'0006	'0006	'0006	'0006	'0006	'0005	'0005	'0005
-3'1	'0010	'0009	'0009	'0009	'0008	'0008	'0008	'0008	'0007	'0007
-3'0	'0014	'0013	'0013	'0012	'0012	'0011	'0011	'0011	'0010	'0010
-2'9	'0019	'0018	'0017	'0017	'0016	'0016	'0015	'0015	'0014	'0014
-2'8	'0026	'0025	'0024	'0023	'0023	'0022	'0021	'0021	'0020	'0019
-2'7	'0035	'0034	'0033	'0032	'0031	'0030	'0029	'0028	'0027	'0026
-2'6	'0047	'0045	'0044	'0043	'0041	'0040	'0039	'0038	'0037	'0036
-2'5	'0062	'0060	'0059	'0057	'0055	'0054	'0052	'0051	'0049	'0048
-2'4	'0082	'0080	'0078	'0075	'0073	'0071	'0069	'0068	'0066	'0064
-2'3	'0107	'0104	'0102	'0099	'0096	'0094	'0091	'0089	'0087	'0084
-2'2	'0139	'0136	'0132	'0129	'0126	'0122	'0119	'0116	'0113	'0110
-2'1	'0179	'0174	'0170	'0166	'0162	'0158	'0154	'0150	'0146	'0143
-2'0	'0228	'0222	'0217	'0212	'0207	'0202	'0197	'0192	'0188	'0183
-1'9	'0287	'0281	'0274	'0268	'0262	'0256	'0250	'0244	'0238	'0233
-1'8	'0359	'0352	'0344	'0336	'0329	'0322	'0314	'0307	'0300	'0294
-1'7	'0446	'0436	'0427	'0418	'0409	'0401	'0392	'0384	'0375	'0367
-1'6	'0548	'0537	'0526	'0516	'0505	'0495	'0485	'0475	'0465	'0455
-1'5	'0668	'0655	'0643	'0630	'0618	'0606	'0594	'0582	'0570	'0559
-1'4	'0808	'0793	'0778	'0764	'0749	'0735	'0722	'0708	'0694	'0681
-1'3	'0968	'0951	'0934	'0918	'0901	'0885	'0869	'0853	'0838	'0823
-1'2	'1151	'1131	'1112	'1093	'1075	'1056	'1038	'1020	'1003	'0985
-1'1	'1357	'1335	'1314	'1292	'1271	'1251	'1230	'1210	'1190	'1170
-1'0	'1587	'1562	'1539	'1515	'1492	'1469	'1446	'1423	'1401	'1379
-0'9	'1841	'1814	'1788	'1762	'1736	'1711	'1685	'1660	'1635	'1611
-0'8	'2119	'2090	'2061	'2033	'2005	'1977	'1949	'1922	'1894	'1867
-0'7	'2420	'2389	'2358	'2327	'2297	'2266	'2236	'2206	'2177	'2148
-0'6	'2743	'2709	'2676	'2643	'2611	'2578	'2546	'2514	'2483	'2451
-0'5	'3085	'3050	'3015	'2981	'2946	'2912	'2877	'2843	'2810	'2776
-0'4	'3446	'3409	'3372	'3336	'3300	'3264	'3228	'3192	'3156	'3121
-0'3	'3821	'3783	'3745	'3707	'3669	'3632	'3594	'3557	'3520	'3483
-0'2	'4207	'4168	'4129	'4090	'4052	'4013	'3974	'3936	'3897	'3859
-0'1	'4620	'4562	'4522	'4483	'4443	'4404	'4364	'4325	'4268	'4247
-0'0	'5000	'4960	'4920	'4880	'4840	'4801	'4761	4721	'4681	'4641

(continúa)

Los valores interiores indican probabilidades. Delante de la coma decimal, (’), se entiende que va un cero. Así, por ejemplo, ’1292 equivale a 0’1292 e indica que $P(Z \leq -1'13) = 0'1292$.

TABLA A. Distribución normal. $P(Z \leq z)$. *(continuación)*

z	0	1	2	3	4	5	6	7	8	9
0'0	'5000	'5040	'5080	'5120	'5160	'5199	'5239	'5279	'5319	'5359
0'1	'5398	'5438	'5478	'5517	'5557	'5596	'5636	'5675	'5714	'5753
0'2	'5793	'5832	'5871	'5910	'5948	'5987	'6026	'6064	'6113	'6141
0'3	'6179	'6217	'6255	'6293	'6331	'6368	'6406	'6443	'6480	'6517
0'4	'6554	'6591	'6628	'6664	'6700	'6736	'6772	'6808	'6844	'6879
0'5	'6915	'6950	'6985	'7019	'7054	'7088	'7123	'7157	'7190	'7224
0'6	'7257	'7291	'7324	'7357	'7389	'7422	'7454	'7486	'7517	'7549
0'7	'7580	'7611	'7642	'7673	'7703	'7734	'7764	'7794	'7823	'7852
0'8	'7881	'7910	'7939	'7967	'7995	'8023	'8051	'8078	'8106	'8133
0'9	'8159	'8186	'8212	'8238	'8264	'8289	'8315	'8340	'8365	'8389
1'0	'8413	'8438	'8461	'8485	'8508	'8531	'8554	'8577	'8599	'8621
1'1	'8643	'8665	'8686	'8708	'8729	'8749	'8770	'8790	'8810	'8830
1'2	'8849	'8869	'8888	'8907	'8925	'8944	'8962	'8980	'8997	'9015
1'3	'9032	'9049	'9066	'9082	'9099	'9115	'9131	'9147	'9162	'9177
1'4	'9192	'9207	'9222	'9236	'9251	'9265	'9278	'9292	'9306	'9319
1'5	'9332	'9345	'9357	'9370	'9382	'9394	'9406	'9418	'9430	'9441
1'6	'9452	'9463	'9474	'9484	'9495	'9505	'9515	'9525	'9535	'9545
1'7	'9554	'9564	'9573	'9582	'9591	'9599	'9608	'9616	'9625	'9633
1'8	'9641	'9648	'9656	'9664	'9671	'9678	'9686	'9693	'9700	'9706
1'9	'9713	'9719	'9726	'9732	'9738	'9744	'9750	'9756	'9762	'9767
2'0	'9772	'9778	'9783	'9788	'9793	'9798	'9803	'9808	'9812	'9817
2'1	'9821	'9826	'9830	'9834	'9838	'9842	'9846	'9850	'9854	'9857
2'2	'9861	'9864	'9868	'9871	'9874	'9878	'9881	'9884	'9887	'9890
2'3	'9893	'9896	'9898	'9901	'9904	'9906	'9909	'9911	'9913	'9916
2'4	'9918	'9920	'9922	'9925	'9927	'9929	'9931	'9932	'9934	'9936
2'5	'9938	'9940	'9941	'9943	'9945	'9946	'9948	'9949	'9951	'9952
2'6	'9953	'9955	'9956	'9957	'9959	'9960	'9961	'9962	'9963	'9964
2'7	'9965	'9966	'9967	'9968	'9969	'9970	'9971	'9972	'9973	'9974
2'8	'9974	'9975	'9976	'9977	'9977	'9978	'9979	'9979	'9980	'9981
2'9	'9981	'9982	'9982	'9983	'9984	'9984	'9985	'9985	'9986	'9986
3'0	'9986	'9987	'9987	'9988	'9988	'9989	'9989	'9989	'9990	'9990
3'1	'9990	'9991	'9991	'9991	'9992	'9992	'9992	'9992	'9993	'9993
3'2	'9993	'9993	'9994	'9994	'9994	'9994	'9994	'9995	'9995	'9995
3'3	'9995	'9995	'9995	'9996	'9996	'9996	'9996	'9996	'9996	'9996
3'4	'9997	'9997	'9997	'9997	'9997	'9997	'9997	'9997	'9997	'9998
3'5	'9998	'9998	'9998	'9998	'9998	'9998	'9998	'9998	'9998	'9998

Los valores interiores indican probabilidades. Delante de la coma decimal, (’), se entiende que va un cero. Así, por ejemplo, ’8925 equivale a 0’8925 e indica que $P(Z < -1'24) = 0'8925$.

FUENTE: BLUM, J. R. y ROSEMBLATT, J. I., *Probabilities and Statistics,* Filadelfia, Launders, 1972.

TABLA B. Distribución χ^2, $P(X \leq \chi^2_{p,n})$.

n	0'005	0'010	0'025	0'050	0'100	0'900	0'950	0'975	0'990	0'995
1	0'000	0'000	0'001	0'004	0'016	2'71	3'84	5'02	6'63	7'88
2	0'010	0'020	0'051	0'103	0'211	4'61	5'99	7'38	9'21	10'60
3	0'072	0'115	0'216	0'352	0'584	6'25	7'81	9'35	11'34	12'84
4	0'207	0'297	0'484	0'711	1'064	7'78	9'49	11'14	13.28	14'86
5	0'412	0'554	0'831	1'145	1'61	9'24	11'07	12'83	15'09	16'75
6	0'68	0'87	1'24	1'64	2'20	10'64	12'59	14'45	16'81	18'55
7	0'99	1'24	1'69	2'17	2'83	12'02	14'07	16'01	18'48	20'28
8	1'34	1'65	2'18	2'73	3'49	13'36	15'51	17'53	20'09	21'96
9	1'73	2'09	2'70	3'33	4'17	14'68	16'92	19'02	21'67	23'59
10	2'16	2'56	3'25	3'94	4'87	15'99	18'31	20'48	23'21	25'19
11	2'60	3'05	3'82	4'57	5'58	17'28	19'68	21'92	24'72	26'76
12	3'07	3'57	4'40	5'23	6'30	18'55	21'03	23'34	26'22	28'30
13	3'57	4'11	5'01	5'89	7'04	19'81	22'36	24'74	27'69	29'82
14	4'07	4'66	5'63	6'57	7'79	21'06	23'68	26'12	29'14	31'32
15	4'60	5'23	6'26	7'26	8'55	22'31	25'00	27'49	30'58	32'80
16	5'14	5'81	6'91	7'96	9'31	23'54	26'30	28'85	32'00	34'27
17	5'70	6'41	7'56	8'67	10'09	24'77	27'59	30'19	33'41	35'72
18	6'26	7'01	8'23	9'39	10'86	25'99	28'87	31'53	34'81	37'16
19	6'84	7'63	8'91	10'12	11'65	27'20	30'14	32'85	36'19	38'58
20	7'43	8'26	8'59	10'85	12'44	28'41	31'41	34'17	37'57	40'00
21	8'03	8'90	10'28	11'59	13'24	29'62	32'67	35'48	38'93	41'40
22	8'64	9'54	10'98	12'34	14'04	30'81	33'92	36'78	40'29	42'80
23	9'26	10'20	11'69	13'09	14'85	32'01	35'17	38'08	41'64	44'18
24	9'89	10'86	12'40	13'85	15'66	33'20	36'42	39'36	42'98	45'56
25	10'52	11'52	13'12	14'61	16'47	34'38	37'65	40'65	44'31	46'93
26	11'16	12'20	13'84	15'38	17'29	35'56	38'89	41'92	45'64	48'29
27	11'81	12'88	14'57	16'15	18'11	36'74	40'11	43'19	46'96	49'64
28	12'46	13'56	15'31	16'39	18'94	37'92	41'34	44'46	48'28	50'99
29	13'21	14'26	16'05	17'71	19'77	39'09	42'56	45'72	49'59	52'34
30	13'79	14'95	16'79	18'49	20'60	40'26	43'77	46'98	50'89	53'67
40	20'71	22'16	24'43	26 51	29'05	51'80	55'76	59'34	63'69	66'77
50	27'99	29'71	32'36	34'76	37'69	63'17	67'50	71'42	76'15	79'49
60	35'53	37'48	40'48	43'19	46'46	74'40	79'08	83'30	88'38	91'95
70	43'28	45'44	48'76	51'74	55'33	85'53	90'53	95'02	100'42	104'22
80	51'17	53'54	57'15	60'39	64'28	96'58	101'88	106'63	112'33	116'32
90	59'20	61'75	65'65	69'13	73'29	107'56	113'14	118'14	124'12	128'30
100	67'33	70'06	74'22	77'93	82'36	118'50	124'34	129'56	135'81	140'17

Los valores centrales de la tabla son los puntos $\chi^2_{p,n}$ que dejan por debajo de sí un área igual a p, supuesto un número n de grados de libertad. Así, por ejemplo $\chi^2_{0,95;11} = 19{,}68$ significa que, para 11 grados de libertad, la probabilidad de obtener un valor igual o menor que 19,68 vale 0,95. El valor 19,68 es la intersección de la columna encabezada por 0,950 y la fila encabezada por 11.

TABLA C. Distribución t, $P(T \leq t_{p,n})$

n	0'900	0'950	0'975	0'990	0'995	0'999
1	3'078	6'314	12'706	31'821	63'657	318'310
2	1'886	2'920	4'303	6'965	9'925	22'327
3	1'638	2'353	3'182	4'541	5'841	10'214
4	1'533	2'132	2'776	3'747	4'604	7'173
5	1'476	2'015	2'571	3'365	4'032	5'893
6	1'440	1'943	2'447	3'143	3'707	5'208
7	1'415	1'895	2'365	2'998	3'499	4'785
8	1'397	1'860	2'306	2'896	3'355	4'501
9	1'383	1'833	2'262	2'821	3'250	4'297
10	1'372	1'812	2'228	2'764	3'169	4'144
11	1'363	1'796	2'201	2'718	3'106	4'025
12	1'356	1'782	2'179	2'681	3'055	3'930
13	1'350	1'771	2'160	2'650	3'012	3'852
14	1'345	1'761	2'145	2'624	2'977	3'787
15	1'341	1'753	2'131	2'602	2'947	3'733
16	1'337	1'746	2'120	2'583	2'921	3'686
17	1'333	1'740	2'110	2'567	2'898	3'646
18	1'330	1'734	2'101	2'552	2'878	3'611
19	1'328	1'729	2'093	2'539	2'861	3'597
20	1'325	1'725	2'086	2'528	2'845	3'552
21	1'323	1'721	2'080	2'518	2'831	3'527
22	1'321	1'717	2'074	2'508	2'819	3'505
23	1'319	1'714	2'069	2'500	2'807	3'485
24	1'318	1'711	2'064	2'492	2'797	3'467
25	1'316	1'708	2'060	2'485	2'787	3'450
26	1'315	1'706	2'056	2'479	2'779	3'435
27	1'314	1'703	2'052	2'473	2'771	3'421
28	1'313	1'701	2'048	2'467	2'763	3'408
29	1'311	1'699	2'045	2'462	2'756	3'396
30	1'310	1'697	2'042	2'457	2'750	3'385
40	1'303	1'684	2'021	2'423	2'704	3'307
50	1'298	1'676	2'009	2'403	2'678	3'262
60	1'296	1'671	2'000	2'390	2'660	3'232
80	1'292	1'664	1'990	2'374	2'639	3'195
100	1'290	1'660	1'984	2'365	2'626	3'174
120	1'289	1'658	1'980	2'358	2'617	3'160
200	1'286	1'653	1'972	2'345	2'601	3'131
500	1'283	1'648	1'965	2'334	2'586	3'106
∞	1'282	1'645	1'960	2'326	2'576	3'090

Los valores centrales de la tabla son los puntos $t_{p,n}$ que dejan por debajo de sí un área a p, supuesto un número n de grados de libertad. Así, por ejemplo, $t_{0,99:21}$ = 129,1518 significa que para 21 grados de libertad la probabilidad de obtener un valor igual o menor que 2,1518 vale 0,99. El valor 2,1518 es la intersección de la columna encabezada por 0,99 y la fila encabezada por 21.

TABLA D

Distribución F, $P(F_{n_1, n_2} \leq f_{0,95,\, n_1, n_2})$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161'4	199'5	215'7	224'6	230'2	234'0	236'8	238'9	240'5	241'9	243'9	245'9	248'0	249'1	250'1	251'1	252'2	253'3	254'3
2	18'51	19'00	19'16	19'25	19'30	19'33	19'35	19'37	19'38	19'40	19'41	19'43	19'45	19'45	19'46	19'47	19'48	19'49	19'50
3	10'13	9'55	9'28	9'12	9'01	8'94	8'89	8'85	8'81	8'79	8'74	8'70	8'66	8'64	8'62	8'59	8'57	8'55	8'53
4	7'71	6'94	6'59	6'39	6'26	6'16	6'09	6'04	6'00	5'96	5'91	5'86	5'80	5'77	5'75	5'72	5'69	5'66	5'63
5	6'61	5'79	5'41	5'19	5'05	4'95	4'88	4'82	4'77	4'74	4'68	4'62	4'56	4'53	4'50	4'46	4'43	4'40	4'36
6	5'99	5'14	4'76	4'53	4'39	4'28	4'21	4'15	4'10	4'06	4'00	3'94	3'87	3'84	3'81	3'77	3'74	3'70	3'67
7	5'59	4'74	4'35	4'12	3'97	3'87	3'79	3'73	3'68	3'64	3'57	3'51	3'44	3'41	3'38	3'34	3'30	3'27	3'23
8	5'32	4'46	4'07	3'84	3'69	3'58	3'50	3'44	3'39	3'35	3'28	3'22	3'15	3'12	3'08	3'04	3'01	2'97	2'93
9	5'12	4'26	3'86	3'63	3'48	3'37	3'29	3'23	3'18	3'14	3'07	3'01	2'94	2'90	2'86	2'83	2'79	2'75	2'71
10	4'96	4'10	3'71	3'48	3'33	3'22	3'14	3'07	3'02	2'98	2'91	2'85	2'77	2'74	2'70	2'66	2'62	2'58	2'54
11	4'84	3'98	3'59	3'36	3'20	3'09	3'01	2'95	2'90	2'85	2'79	2'72	2.65	2'61	2'57	2'53	2'49	2'45	2'40
12	4'75	3'89	3'49	3'26	3'11	3'00	2'91	2'85	2'80	2'75	2'69	2'62	2'54	2'51	2'47	2'43	2'38	2'34	2'30
13	4'67	3'81	3'41	3'18	3'03	2'92	2'83	2'77	2'71	2'67	2'60	2'53	2'46	2'42	2'38	2'34	2'30	2'25	2'21
14	4'60	3'74	3'34	3'11	2'96	2'85	2'76	2'70	2'65	2'60	2'53	2'46	2'39	2'35	2'31	2'27	2'22	2'18	2'13
15	4'54	3'68	3'29	3'06	2'90	2'79	2'71	2'64	2'59	2'54	2'48	2'40	2'33	2'29	2'25	2'20	2'16	2'11	2'07
16	4'49	3'63	3'24	3'01	2'85	2'74	2'66	2'59	2'54	2'49	2'42	2'35	2'28	2'24	2'19	2'15	2'11	2'06	2'01
17	4'45	3'59	3'20	2'96	2'81	2'70	2'61	2'55	2'49	2'45	2'38	2'31	2'23	2'19	2'15	2'10	2'06	2'01	1'96
18	4'41	3'55	3'16	2'93	2'77	2'66	2'58	2'51	2'46	2'41	2'34	2'27	2'19	2'15	2'11	2'06	2'02	1'97	1'92
19	4'38	3'52	3'13	2'90	2'74	2'63	2'54	2'48	2'42	2'38	2'31	2'23	2'16	2'11	2'07	2'03	1'98	1'93	1'88
20	4'35	3'49	3'10	2'87	2'71	2'60	2'51	2'45	2'39	2'35	2'28	2'20	2'12	2'08	2'04	1'99	1'95	1'90	1'84
21	4'32	3'47	3'07	2'84	2'68	2'57	2'49	2'42	2'37	2'32	2'25	2'18	2'10	2'05	2'01	1'96	1'92	1'87	1'81
22	4'30	3'44	3'05	2'82	2'66	2'55	2'46	2'40	2'34	2'30	2'23	2'15	2'07	2'03	1'98	1'94	1'89	1'84	1'78
23	4'28	3'42	3'03	2'80	2'64	2'53	2'44	2'37	2'32	2'27	2'20	2'13	2'05	2'01	1'96	1'91	1'86	1'81	1'76
24	4'26	3'40	3'01	2'78	2'62	2'51	2'41	2'36	2'30	2'25	2'18	2'11	2'03	1'98	1'94	1'89	1'84	1'79	1'73
25	4'24	3'39	2'99	2'76	2'60	2'49	2'40	2'34	2'28	2'24	2'16	2'09	2'01	1'96	1'92	1'87	1'82	1'77	1'71
26	4'23	3'37	2'98	2'74	2'59	2'47	2'39	2'32	2'27	2'22	2'15	2'07	1'99	1'95	1'90	1'85	1'80	1'75	1'69
27	4'21	3'35	2'96	2'73	2'57	2'46	2'37	2'31	2'25	2'20	2'13	2'06	1'97	1'93	1'88	1'84	1'79	1'73	1'67
28	4'20	3'34	2'95	2'71	2'56	2'45	2'36	2'29	2'24	2'19	2'12	2'04	1'96	1'91	1'87	1'82	1'77	1'71	1'65
29	4'18	3'33	2'93	2'70	2'55	2'43	2'35	2'28	2'22	2'18	2'10	2'03	1'94	1'90	1'85	1'81	1'75	1'70	1'64
30	4'17	3'32	2'92	2'69	2'53	2'42	2'33	2'27	2'21	2'16	2'09	2'01	1'93	1'89	1'84	1'79	1'74	1'68	1'62
40	4'08	3'23	2'84	2'61	2'45	2'34	2'25	2'18	2'12	2'08	2'00	1'92	1'84	1'79	1'74	1'69	1'64	1'58	1'51
60	4'00	3'15	2'76	2'53	2'37	2'25	2'17	2'10	2'04	1'99	1'92	1'84	1'75	1'70	1'65	1'59	1'53	1'47	1'39
120	3'92	3'07	2'68	2'45	2'29	2'17	2'09	2'02	1'96	1'91	1'83	1'75	1'66	1'61	1'55	1'50	1'43	1'35	1'25
∞	3'84	3'00	2'60	2'37	2'21	2'10	2'01	1'94	1'88	1'83	1'75	1'67	1'57	1'52	1'46	1'39	1'32	1'22	1'00

TABLA D

Distribución F, $P(F_{n_1, n_2} \leq f_{0,975, n_1, n_2})$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647'8	799'5	864'2	899'6	921'8	937'1	948'2	956'7	963'3	968'6	976'7	948'9	993'1	997'2	1.001	1.006	1.010	1.014	1.018
2	38'51	39'00	39'17	39'25	39'30	39'33	39'36	39'37	39'39	39'40	39'41	39'43	39'45	39'46	39'46	39'47	39'48	39'49	39'50
3	17'44	16'04	15'44	15'10	14'88	14'73	14'62	14'54	14'47	14'42	14'34	14'25	14'17	14'12	14'08	14'04	13'99	13'95	13'90
4	12'22	10'65	9'98	9'60	9'36	9'20	9'07	8'98	8'90	8'84	8'75	8'66	8'56	8'51	8'46	8'41	8'36	8'31	8'26
5	10'01	8'43	7'76	7'39	7'15	6'98	6'85	6'76	6'68	6'62	6'52	6'43	6'33	6'28	6'23	6'18	6'12	6'07	6'02
6	8'81	7'26	6'60	6'23	5'99	5'82	5'70	5'60	5'52	5'46	5'37	5'27	5'17	5'12	5'07	5'01	4'96	4'90	4'85
7	8'07	6'54	5'89	5'52	5'29	5'12	4'99	4'90	4'82	4'76	4'67	4'57	4'47	4'42	4'36	4'31	4'25	4'20	4'14
8	7'57	6'06	5'42	5'05	4'82	4'65	4'53	4'43	4'36	4'30	4'20	4'10	4'00	3'95	3'89	3'84	3'78	3'73	3'67
9	7'21	5'71	5'08	4'72	4'48	4'32	4'20	4'10	4'03	3'96	3'87	3'77	3'67	3'61	3'56	3'51	3'45	3'39	3'33
10	6'94	5'46	4'83	4'47	4'24	4'07	3'95	3'85	3'78	3'72	3'62	3'52	3'42	3'37	3'31	3'26	3'20	3'14	3'08
11	6'72	5'26	4'63	4'28	4'04	3'88	3'76	3'66	3'59	3'53	3'43	3'33	3'23	3'17	3'12	3'06	3'00	2'94	2'88
12	6'55	5'10	4'47	4'12	3'89	3'73	3'61	3'51	3'44	3'37	3'28	3'18	3'07	3'02	2'96	2'91	2'85	2'79	2'72
13	6'41	4'97	4'35	4'00	3'77	3'60	3'48	3'39	3'31	3'25	3'15	3'05	2'95	2'89	2'84	2'78	2'72	2'66	2'60
14	6'30	4'86	4'24	3'89	3'66	3'50	3'38	3'29	3'21	3'15	3'05	2'95	2'84	2'79	2'73	2'67	2'61	2'55	2'49
15	6'20	4'77	4'15	3'80	3'58	3'41	3'29	3'20	3'12	3'06	2'96	2'86	2'76	2'70	2'64	2'59	2'52	2'46	2'40
16	6'12	4'69	4'08	3'73	3'50	3'34	3'22	3'12	3'05	2'99	2'89	2'79	2'68	2'63	2'57	2'51	2'45	2'38	2'32
17	6'04	4'62	4'01	3'66	3'44	3'28	3'16	3'06	2'98	2'92	2'82	2'72	2'62	2'56	2'50	2'44	2'38	2'32	2'25
18	5'98	4'56	3'95	3'61	3'38	3'22	3'10	3'01	2'93	2'87	2'77	2'67	2'56	2'50	2'44	2'38	2'32	2'26	2'19
19	5'92	4'51	3'90	3'56	3'33	3'17	3'05	2'96	2'88	2'82	2'72	2'62	2'51	2'45	2'39	2'33	2'27	2'20	2'13
20	5'87	4'46	3'86	3'51	3'29	3'13	3'01	2'91	2'84	2'77	2'68	2'57	2'46	2'41	2'35	2'29	2'22	2'16	2'09
21	5'83	4'42	3'82	3'48	3'25	2'09	2'97	2'87	2'80	2'73	2'64	2'53	2'42	2'37	2'31	2'25	2'18	2'11	2'04
22	5'79	4'38	3'78	3'44	3'22	3'05	2'93	2'84	2'76	2'70	2'60	2'50	2'39	2'33	2'27	2'21	2'14	2'08	2'00
23	5'75	4'35	3'75	3'41	3'18	3'02	2'90	2'81	2'73	2'67	2'57	2'47	2'36	2'30	2'24	2'18	2'11	2'04	1'97
24	5'72	4'32	3'72	3'38	3'15	2'99	2'87	2'78	2'70	2'64	2'54	2'44	2'33	2'27	2'21	2'15	2'08	2'01	1'94
25	5'69	4'29	3'69	3'35	3'13	2'97	2'85	2'75	2'68	2'61	2'51	2'41	2'30	2'24	2'18	2'12	2'05	1'98	1'91
26	5'66	4'27	3'67	3'33	3'10	2'94	2'82	2'73	2'65	2'59	2'49	2'39	2'28	2'22	2'16	2'09	2'03	1'95	1'88
27	5'63	4'24	3'65	3'31	3'08	2'92	2'80	2'71	2'63	2'57	2'47	2'36	2'25	2'19	2'13	2'07	2'00	1'93	1'85
28	5'61	4'22	3'63	3'29	3'06	2'90	2'78	2'69	2'61	2'55	2'45	2'34	2'23	2'17	2'11	2'05	1'98	1'91	1'83
29	5'59	4'20	3'61	3'27	3'04	2'88	2'76	2'67	2'59	2'53	2'43	2'32	2'21	2'15	2'09	2'03	1'96	1'89	1'81
30	5'57	4'18	3'59	3'25	3'03	2'87	2'75	2'65	2'57	2'51	2'41	2'31	2'20	2'14	2'07	2'01	1'94	1'87	1'79
40	5'42	4'05	3'46	3'13	2'90	2'74	2'62	2'53	2'45	2'39	2'29	2'18	2'07	2'01	1'94	1'88	1'80	1'72	1'64
60	5'29	3'93	3'34	3'01	2'79	2'63	2'51	2'41	2'33	2'27	2'17	2'06	1'94	1'88	1'82	1'74	1'67	1'58	1'48
120	5'15	3'80	3'23	2'89	2'67	2'52	2'39	2'30	2'22	2'16	2'05	1'94	1'82	1'76	1'69	1'61	1'53	1'43	1'31
∞	5'02	3'69	3'12	2'79	2'59	2'41	2'29	2'19	2'11	2'05	1'94	1'83	1'71	1'64	1'57	1'48	1'39	1'27	1'00

TABLA D

Distribución F, $P(F_{n_1, n_2} \leq f_{0{,}990,\, n_1, n_2})$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98'50	99'00	99'17	99'25	99'30	99'33	99'36	99'37	99'39	99'40	99'42	99'43	99'45	99'46	99'47	99'47	99'48	99'49	99'50
3	34'12	30'82	29'46	28'71	28'24	27'91	27'67	27'49	27'35	27'23	27'05	26'87	26'69	26'60	26'50	26'41	26'32	26'22	26'13
4	21'20	18'00	16'69	15'98	15'52	15'21	14'98	14'80	14'66	14'55	14'37	14'20	14'02	13'93	13'84	13'75	13'65	13'56	13'46
5	16'26	13'27	12'06	11'39	10'97	10'67	10'46	10'29	10'16	10'05	9'89	9'72	9'55	9'47	9'38	9'29	9'20	9'11	9'02
6	13'75	10'92	9'78	9'15	8'75	8'47	8'26	8'10	7'98	7'87	7'72	7'56	7'40	7'31	7'23	7'14	7'06	6'97	6'88
7	12'25	9'55	8'45	7'85	7'46	7'19	6'99	6'84	6'72	6'62	6'47	6'31	6'16	6'07	5'99	5'91	5'82	5'74	5'65
8	11'26	8'65	7'59	7'01	6'63	6'37	6'18	6'03	5'91	5'81	5'67	5'52	5'36	5'28	5'20	5'12	5'03	4'95	4'86
9	10'56	8'02	6'99	6'42	6'06	5'80	5'61	5'47	5'35	5'26	5'11	4'96	4'91	4'73	4'65	4'57	4'48	4'40	4'31
10	10'04	7'56	6'55	5'99	5'64	5'39	5'20	5'06	4'94	4'85	4'71	4'56	4'41	4'33	4'25	4'17	4'08	4'00	3'91
11	9'65	7'21	6'22	5'67	5'32	5'07	4'89	4'74	4'63	4'54	4'40	4'25	4'10	4'02	3'94	3'86	3'78	3'69	3'60
12	9'33	6'93	5'95	5'41	5'06	4'82	4'64	4'50	4'39	4'30	4'16	4'01	3'86	3'78	3'70	3'62	3'54	3'45	3'36
13	9'07	6'70	5'74	5'21	4'86	4'62	4'44	4'30	4'19	4'10	3'96	3'82	3'66	3'59	3'51	3'43	3'34	3'25	3'17
14	8'86	6'51	5'56	5'04	4'69	4'46	4'28	4'14	4'03	3'94	3'80	3'66	3'51	3'43	3'35	3'27	3'18	3'09	3'00
15	8'68	6'36	5'42	4'89	4'56	4'32	4'14	4'00	3'89	3'80	3'67	3'52	3'37	3'29	3'21	3'13	3'05	2'96	2'87
16	8'53	6'23	5'29	4'77	4'44	4'20	4'03	3'89	3'78	3'69	3'55	3'41	3'26	3'18	3'10	3'02	2'93	2'84	2'75
17	8'40	6'11	5'18	4'67	4'34	4'10	3'93	3'79	3'68	3'59	3'46	3'31	3'16	3'08	3'00	2'92	2'83	2'75	2'65
18	8'29	6'01	5'09	4'58	4'25	4'01	3'84	3'71	3'60	3'51	3'37	3'23	3'08	3'00	2'92	2'84	2'75	2'66	2'57
19	8'18	5'93	5'01	4'50	4'17	3'94	3'77	3'63	3'52	3'43	3'30	3'15	3'00	2'92	2'84	2'76	2'67	2'58	2'49
20	8'10	5'85	4'94	4'43	4'10	3'87	3'70	3'56	3'46	3'37	3'23	3'09	2'94	2'86	2'78	2'69	2'61	2'52	2'42
21	8'02	5'78	4'87	4'37	4'04	3'81	3'64	3'51	3'40	3'31	3'17	3'03	2'88	2'80	2'72	2'64	2'55	2'46	2'36
22	7'95	5'72	4'82	4'31	3'99	3'76	3'59	3'45	3'35	3'26	3'12	2'98	2'83	2'75	2'67	2'58	2'50	2'40	2'31
23	7'88	5'66	4'76	4'26	3'94	3'71	3'54	3'41	3'30	3'21	3'07	2'93	2'78	2'70	2'62	2'54	2'45	2'35	2'26
24	7'82	5'61	4'72	4'22	3'90	3'67	3'50	3'36	3'26	3'17	3'03	2'89	2'74	2'66	2'58	2'49	2'40	2'31	2'21
25	7'77	5'57	4'68	4'18	3'85	3'63	3'46	3'32	3'22	3'13	2'99	2'85	2'70	2'62	2'54	2'45	2'36	2'27	2'17
26	7'72	5'53	4'64	4'14	3'82	3'59	3'42	3'29	3'18	3'09	2'96	2'81	2'66	2'58	2'50	2'42	2'33	2'23	2'13
27	7'68	5'49	4'60	4'11	3'78	3'56	3'39	3'26	3'15	3'06	2'93	2'78	2'63	2'55	2'47	2'38	2'29	2'20	2'10
28	7'64	5'45	4'57	4'07	3'75	3'53	3'36	3'23	3'12	3'03	2'90	2'75	2'60	2'52	2'44	2'35	2'26	2'17	2'06
29	7'60	5'42	4'54	4'04	3'73	3'50	3'33	3'20	3'09	3'00	2'87	2'63	2'57	2'49	2'41	2'33	2'23	2'14	2'03
30	7'56	5'39	4'51	4'02	3'70	3'47	3'30	3'17	3'07	2'98	2'84	2'70	2'55	2'47	2'39	2'30	2'21	2'11	2'01
40	7'31	5'18	4'31	3'83	3'51	3'29	3'12	2'99	2'89	2'80	2'66	2'52	2'37	2'29	2'20	2'11	2'02	1'92	1'80
60	7'08	4'98	4'13	3'65	3'34	3'12	2'95	2'82	2'72	2'63	2'50	2'35	2'20	2'12	2'03	1'94	1'84	1'73	1'60
120	6'85	4'79	3'95	3'48	3'17	2'96	2'79	2'66	2'56	2'47	2'34	2'19	2'03	1'95	1'86	1'76	1'66	1'53	1'38
∞	6'63	4'61	3'78	3'32	3'02	2'80	2'64	2'51	2'41	2'32	2'18	2'04	1'88	1'79	1'70	1'59	1'47	1'32	1'00

TABLA D

Distribución F, $P(F_{n_1, n_2} \leq f_{0,995, n_1, n_2})$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	24940	25044	25148	25253	25359	25465
2	198'5	199'0	199'2	199'2	199'3	199'3	199'4	199'4	199'4	199'4	199'4	199'4	199'4	199'5	199'5	199'5	199'5	199'5	199'5
3	55'55	49'80	47'47	46'19	45'39	44'84	44'43	44'13	43'88	43'69	43'39	43'08	42'78	42'62	42'47	42'31	42'15	41'99	41'83
4	31'33	26'28	24'26	23'15	22'46	21'97	21'62	21'35	21'14	20'97	20'70	20'44	20'17	20'03	19'89	19'75	19'61	19'47	19'32
5	22'72	18'31	16'53	15'56	14'94	14'51	14'20	13'96	13'77	13'62	13'38	13'15	12'90	12'78	12'66	12'53	12'40	12'27	12'14
6	18'63	14'54	12'92	12'03	11'46	11'07	10'79	10'57	10'39	10'25	10'03	9'81	9'59	9'47	9'36	9'24	9'12	9'00	8'88
7	16'24	12'40	10'88	10'05	9'52	9'16	8'89	8'68	8'51	8'38	8'18	7'97	7'75	7'65	7'53	7'42	7'31	7'19	7'08
8	14'69	11'04	9'61	8'81	8'30	7'95	7'69	7'50	7'34	7'21	7'01	6'81	6'61	6'50	6'40	6'29	6'18	6'06	5'95
9	13'61	10'11	8'72	7'96	7'47	7'13	6'88	6'69	6'54	6'42	6'23	6'03	5'83	5'73	5'62	5'52	5'41	5'30	5'19
10	12'83	9'43	8'08	7'34	6'87	6'54	6'30	6'12	5'97	5'85	5'66	5'47	5'27	5'17	5'07	4'97	4'86	4'75	4'64
11	12'23	8'91	7'60	6'88	6'42	6'10	5'86	5'68	5'54	5'42	5'24	5'05	4'86	4'76	4'65	4'55	4'44	4'34	4'23
12	11'75	8'51	7'23	6'52	6'07	5'76	5'52	5'35	5'20	5'09	4'91	4'72	4'53	4'43	4'33	4'23	4'12	4'01	3'90
13	11'37	8'19	6'93	6'23	5'79	5'48	5'25	5'08	4'94	4'82	4'64	4'46	4'27	4'17	4'07	3'97	3'87	3'76	3'65
14	11'06	7'92	6'68	6'00	5'56	5'26	5'03	4'86	4'72	4'60	4'43	4'25	4'06	3'96	3'86	3'76	3'66	3'55	3'44
15	10'80	7'70	6'48	5'80	5'37	5'07	4'85	4'67	4'54	4'42	4'25	4'07	3'88	3'79	3'69	3'58	3'48	3'37	3'26
16	10'58	7'51	6'30	5'64	5'21	4'91	4'69	4'52	4'38	4'27	4'10	3'92	3'73	3'64	3'54	3'44	3'33	3'22	3'11
17	10'38	7'35	6'16	5'50	5'07	4'78	4'56	4'39	4'25	4'14	3'97	3'79	3'61	3'51	3'41	3'31	3'21	3'10	2'98
18	10'22	7'21	6'03	5'37	4'96	4'66	4'44	4'28	4'14	4'03	3'86	3'68	3'50	3'40	3'30	3'20	3'10	2'99	2'87
19	10'07	7'09	5'92	5'27	4'85	4'56	4'34	4'18	4'04	3'93	3'76	3'59	3'40	3'31	3'21	3'11	3'00	2'89	2'78
20	9'94	6'69	5'82	5'17	4'76	4'47	4'26	4'09	3'96	3'85	3'68	3'50	3'32	3'22	3'12	3'02	2'92	2'81	2'69
21	9'83	6'89	5'73	5'09	4'68	4'39	4'18	4'01	3'88	3'77	3'60	3'43	3'24	3'15	3'05	2'95	2'84	2'73	2'61
22	9'73	6'81	5'65	5'02	4'61	4'32	4'11	3'94	3'81	3'70	3'54	3'36	3'18	3'08	2'98	2'88	2'77	2'66	2'55
23	9'63	6'73	5'58	4'95	4'54	4'26	4'05	3'88	3'75	3'64	3'57	3'30	3'12	3'02	2'92	2'82	2'71	2'60	2'48
24	9'55	6'66	5'52	4'89	4'49	4'20	3'99	3'83	3'69	3'59	3'42	3'25	3'06	2'97	2'87	2'77	2'66	2'55	2'43
25	9'48	6'60	5'46	4'84	4'43	4'15	3'94	3'78	3'64	3'54	3'37	3'20	3'01	2'92	2'82	2'72	2'61	2'50	2'38
26	9'41	6'54	5'41	4'79	4'38	4'10	3'89	3'73	3'60	3'49	3'33	3'15	2'97	2'87	2'77	2'67	2'56	2'45	2'33
27	9'34	6'49	5'36	4'74	4'34	4'06	3'85	3'69	3'56	3'45	3'28	3'11	2'93	2'83	2'73	2'63	2'52	2'41	2'29
28	9'28	6'44	5'32	4'70	4'30	4'02	3'81	3'65	3'52	3'41	3'25	3'07	2'89	2'79	2'69	2'59	2'48	2'37	2'25
29	9'23	6'40	5'28	4'66	4'26	3'98	3'77	3'61	3'48	3'38	3'21	3'04	2'86	2'76	2'66	2'56	2'45	2'33	2'21
30	9'18	6'35	5'24	4'62	4'23	3'95	3'74	3'58	3'45	3'34	3'18	3'01	2'82	2'73	2'63	2'52	2'42	2'30	2'18
40	8'83	6'07	4'98	4'37	3'99	3'71	3'51	3'35	3'22	3'12	2'95	2'78	2'60	2'50	2'40	2'30	2'18	2'06	1'93
60	8'49	5'79	4'73	4'14	3'76	3'49	3'29	3'13	3'01	2'90	2'74	2'57	2'39	2'29	2'19	2'08	1'96	1'83	1'69
120	8'18	5'54	4'50	3'92	3'55	3'28	3'09	2'93	2'81	2'71	2'54	2'37	2'19	2'09	1'98	1'87	1'75	1'61	1'43
∞	7'88	5'30	4'28	3'72	3'35	3'09	2'90	2'74	2'62	2'52	2'36	2'19	2'00	1'90	1'79	1'67	1'53	1'36	1'00

TABLA E

Distribución binomial, $P(X \leq k)$.

		p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	k											
$n=1$	0	.950	.900	.800	.700	.600	.500	.400	.300	.200	.100	.050
	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=2$	0	.902	.810	.640	.490	.360	.250	.160	.090	.040	.010	.002
	1	.997	.990	.960	.910	.840	.750	.640	.510	.360	.190	.097
	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=3$	0	.857	.729	.512	.343	.216	.125	.064	.027	.008	.001	.000
	1	.993	.972	.896	.784	.648	.500	.352	.216	.104	.028	.007
	2	1.000	.999	.992	.973	.936	.875	.784	.657	.488	.271	.143
	3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=4$	0	.815	.656	.410	.240	.130	.063	.026	.008	.002	.000	.000
	1	.986	.948	.819	.652	.475	.313	.179	.084	.027	.004	.000
	2	1.000	.996	.973	.916	.821	.688	.525	.348	.181	.052	.014
	3	1.000	1.000	.998	.992	.974	.938	.870	.760	.590	.344	.185
	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=5$	0	.774	.590	.328	.168	.078	.031	.010	.002	.000	.000	.000
	1	.977	.919	.737	.528	.337	.188	.087	.031	.007	.000	.000
	2	.999	.991	.942	.837	.683	.500	.317	.163	.058	.009	.001
	3	1.000	1.000	.993	.969	.913	.813	.663	.472	.263	.081	.023
	4	1.000	1.000	1.000	.998	.990	.969	.922	.832	.672	.410	.226
	5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=6$	0	.735	.531	.262	.118	.047	.016	.004	.001	.000	.000	.000
	1	.967	.886	.655	.420	.233	.109	.041	.011	.002	.000	.000
	2	.998	.984	.901	.744	.544	.344	.179	.070	.017	.001	.000
	3	1.000	.999	.983	.930	.821	.656	.456	.256	.099	.016	.002
	4	1.000	1.000	.998	.989	.959	.891	.767	.580	.345	.114	.033
	5	1.000	1.000	1.000	.999	.996	.984	.953	.882	.738	.469	.265
	6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=7$	0	.698	.478	.210	.082	.028	.008	.002	.000	.000	.000	.000
	1	.956	.850	.577	.329	.159	.063	.019	.004	.000	.000	.000
	2	.996	.974	.852	.647	.420	.227	.096	.029	.005	.000	.000
	3	1.000	.997	.967	.874	.710	.500	.290	.126	.033	.003	.000
	4	1.000	1.000	.995	.971	.904	.773	.580	.353	.148	.026	.004

TABLA E

(continuación)

		p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	k											
	5	1.000	1.000	1.000	.996	.981	.938	.841	.671	.423	.150	.044
	6	1.000	1.000	1.000	1.000	.998	.992	.972	.918	.790	.522	.302
	7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=8$	0	.663	.430	.168	.058	.017	.004	.001	.000	.000	.000	.000
	1	.943	.813	.503	.255	.106	.035	.009	.001	.000	.000	.000
	2	.994	.962	.797	.552	.315	.145	.050	.011	.001	.000	.000
	3	1.000	.995	.944	.806	.594	.363	.174	.058	.010	.000	.000
	4	1.000	1.000	.990	.942	.826	.637	.406	.194	.056	.005	.000
	5	1.000	1.000	.999	.989	.950	.855	.685	.448	.203	.038	.006
	6	1.000	1.000	1.000	.999	.991	.965	.894	.745	.497	.187	.057
	7	1.000	1.000	1.000	1.000	.999	.996	.983	.942	.832	.570	.337
	8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=9$	0	.630	.387	.134	.040	.010	.002	.000	.000	.000	.000	.000
	1	.929	.775	.436	.196	.071	.020	.004	.000	.000	.000	.000
	2	.992	.947	.738	.463	.232	.090	.025	.004	.000	.000	.000
	3	.999	.992	.914	.730	.483	.254	.099	.025	.003	.000	.000
	4	1.000	.999	.980	.901	.733	.500	.267	.099	.020	.001	.000
	5	1.000	1.000	.997	.975	.901	.746	.517	.270	.086	.008	.001
	6	1.000	1.000	1.000	.996	.975	.910	.768	.537	.262	.053	.008
	7	1.000	1.000	1.000	1.000	.996	.980	.929	.804	.564	.225	.071
	8	1.000	1.000	1.000	1.000	1.000	.998	.990	.960	.866	.613	.370
	9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=10$	0	.599	.349	.107	.028	.006	.001	.000	.000	.000	.000	.000
	1	.914	.736	.376	.149	.046	.011	.002	.000	.000	.000	.000
	2	.988	.930	.678	.383	.167	.055	.012	.002	.000	.000	.000
	3	.999	.987	.879	.650	.382	.172	.055	.011	.001	.000	.000
	4	1.000	.998	.967	.850	.633	.377	.166	.047	.006	.000	.000
	5	1.000	1.000	.994	.953	.834	.623	.367	.150	.033	.002	.000
	6	1.000	1.000	.999	.989	.945	.828	.618	.350	.121	.013	.001
	7	1.000	1.000	1.000	.998	.988	.945	.833	.617	.322	.070	.012
	8	1.000	1.000	1.000	1.000	.998	.989	.954	.851	.624	.264	.086
	9	1.000	1.000	1.000	1.000	1.000	.999	.994	.972	.893	.651	.401
	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=11$	0	.569	.314	.086	.020	.004	.000	.000	.000	.000	.000	.000
	1	.898	.697	.322	.113	.030	.006	.001	.000	.000	.000	.000
	2	.985	.910	.617	.313	.119	.033	.006	.001	.000	.000	.000
	3	.998	.981	.839	.570	.296	.113	.029	.004	.000	.000	.000

TABLA E

(continuación)

		p									
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	k											
	4	1.000	.997	.950	.790	.533	.274	.099	.022	.002	.000	.000
	5	1.000	1.000	.988	.922	.753	.500	.247	.078	.012	.000	.000
	6	1.000	1.000	.998	.978	.901	.726	.467	.210	.050	.003	.000
	7	1.000	1.000	1.000	.996	.971	.887	.704	.430	.161	.019	.002
	8	1.000	1.000	1.000	.999	.994	.967	.881	.687	.383	.090	.015
	9	1.000	1.000	1.000	1.000	.999	.994	.970	.887	.678	.303	.102
	10	1.000	1.000	1.000	1.000	1.000	1.000	.996	.980	.914	.686	.431
	11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 12$	0	.540	.282	.069	.014	.002	.000	.000	.000	.000	.000	.000
	1	.882	.659	.275	.085	.020	.003	.000	.000	.000	.000	.000
	2	.980	.889	.558	.253	.083	.019	.003	.000	.000	.000	.000
	3	.998	.974	.795	.493	.225	.073	.015	.002	.000	.000	.000
	4	1.000	.996	.927	.724	.438	.194	.057	.009	.001	.000	.000
	5	1.000	.999	.981	.882	.665	.387	.158	.039	.004	.000	.000
	6	1.000	1.000	.996	.961	.842	.613	.335	.118	.019	.001	.000
	7	1.000	1.000	.999	.991	.943	.806	.562	.276	.073	.004	.000
	8	1.000	1.000	1.000	.998	.985	.927	.775	.507	.205	.026	.002
	9	1.000	1.000	1.000	1.000	.997	.981	.917	.747	.442	.111	.020
	10	1.000	1.000	1.000	1.000	1.000	.997	.980	.915	.725	.341	.118
	11	1.000	1.000	1.000	1.000	1.000	1.000	.998	.986	.931	.718	.460
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 13$	0	.513	.254	.055	.010	.001	.000	.000	.000	.000	.000	.000
	1	.865	.621	.234	.064	.013	.002	.000	.000	.000	.000	.000
	2	.975	.866	.502	.202	.058	.011	.001	.000	.000	.000	.000
	3	.997	.966	.747	.421	.169	.046	.008	.001	.000	.000	.000
	4	1.000	.994	.901	.654	.353	.133	.032	.004	.000	.000	.000
	5	1.000	.999	.970	.835	.574	.291	.098	.018	.001	.000	.000
	6	1.000	1.000	.993	.938	.771	.500	.229	.062	.007	.000	.000
	7	1.000	1.000	.999	.982	.902	.709	.426	.165	.030	.001	.000
	8	1.000	1.000	1.000	.996	.968	.867	.647	.346	.099	.006	.000
	9	1.000	1.000	1.000	.999	.992	.954	.831	.579	.253	.034	.003
	10	1.000	1.000	1.000	1.000	.999	.989	.942	.798	.498	.134	.025
	11	1.000	1.000	1.000	1.000	1.000	.998	.987	.936	.766	.379	.135
	12	1.000	1.000	1.000	1.000	1.000	1.000	.999	.990	.945	.746	.487
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 14$	0	.488	.229	.044	.007	.001	.000	.000	.000	.000	.000	.000
	1	.847	.585	.198	.047	.008	.001	.000	.000	.000	.000	.000
	2	.970	.842	.448	.161	.040	.006	.001	.000	.000	.000	.000

TABLA E

(continuación)

		p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	k											
	3	.996	.956	.698	.355	.124	.029	.004	.000	.000	.000	.000
	4	1.000	.991	.870	.584	.279	.090	.018	.002	.000	.000	.000
	5	1.000	.999	.956	.781	.486	.212	.058	.008	.000	.000	.000
	6	1.000	1.000	.988	.907	.692	.395	.150	.031	.002	.000	.000
	7	1.000	1.000	.998	.969	.850	.605	.308	.093	.012	.000	.000
	8	1.000	1.000	1.000	.992	.942	.788	.514	.219	.044	.001	.000
	9	1.000	1.000	1.000	.998	.982	.910	.721	.416	.130	.009	.000
	10	1.000	1.000	1.000	1.000	.996	.971	.876	.645	.302	.044	.004
	11	1.000	1.000	1.000	1.000	.999	.994	.960	.839	.552	.158	.030
	12	1.000	1.000	1.000	1.000	1.000	.999	.992	.953	.802	.415	.153
	13	1.000	1.000	1.000	1.000	1.000	1.000	.999	.993	.956	.771	.512
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 15$	0	.463	.206	.035	.005	.000	.000	.000	.000	.000	.000	.000
	1	.829	.549	.167	.035	.005	.000	.000	.000	.000	.000	.000
	2	.964	.816	.398	.127	.027	.004	.000	.000	.000	.000	.000
	3	.995	.944	.648	.297	.091	.018	.002	.000	.000	.000	.000
	4	.999	.987	.836	.515	.217	.059	.009	.001	.000	.000	.000
	5	1.000	.998	.939	.722	.403	.151	.034	.004	.000	.000	.000
	6	1.000	1.000	.982	.869	.610	.304	.095	.015	.001	.000	.000
	7	1.000	1.000	.996	.950	.787	.500	.213	.050	.004	.000	.000
	8	1.000	1.000	.999	.985	.905	.696	.390	.131	.018	.000	.000
	9	1.000	1.000	1.000	.996	.966	.849	.597	.278	.061	.002	.000
	10	1.000	1.000	1.000	.999	.991	.941	.783	.485	.164	.013	.001
	11	1.000	1.000	1.000	1.000	.998	.982	.909	.703	.352	.056	.005
	12	1.000	1.000	1.000	1.000	1.000	.996	.973	.873	.602	.184	.036
	13	1.000	1.000	1.000	1.000	1.000	1.000	.995	.965	.833	.451	.171
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.965	.794	.537
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 16$	0	.440	.185	.028	.003	.000	.000	.000	.000	.000	.000	.000
	1	.811	.515	.141	.026	.003	.000	.000	.000	.000	.000	.000
	2	.957	.789	.352	.099	.018	.002	.000	.000	.000	.000	.000
	3	.993	.932	.598	.246	.065	.011	.001	.000	.000	.000	.000
	4	.999	.983	.798	.450	.167	.038	.005	.000	.000	.000	.000
	5	1.000	.997	.918	.660	.329	.105	.019	.002	.000	.000	.000
	6	1.000	.999	.973	.825	.527	.227	.058	.007	.000	.000	.000
	7	1.000	1.000	.993	.926	.716	.402	.142	.026	.001	.000	.000
	8	1.000	1.000	.999	.974	.858	.598	.284	.074	.007	.000	.000
	9	1.000	1.000	1.000	.993	.942	.773	.473	.175	.027	.001	.000
	10	1.000	1.000	1.000	.998	.981	.895	.671	.340	.082	.003	.000

TABLA E

(continuación)

		p									
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	k											
	11	1.000	1.000	1.000	1.000	.995	.962	.833	.550	.202	.017	.001
	12	1.000	1.000	1.000	1.000	.999	.989	.935	.754	.402	.068	.007
	13	1.000	1.000	1.000	1.000	1.000	.998	.982	.901	.648	.211	.043
	14	1.000	1.000	1.000	1.000	1.000	1.000	.997	.974	.859	.485	.189
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.997	.972	.815	.560
	16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 17$	0	.418	.167	.023	.002	.000	.000	.000	.000	.000	.000	.000
	1	.792	.482	.118	.019	.002	.000	.000	.000	.000	.000	.000
	2	.950	.762	.310	.077	.012	.001	.000	.000	.000	.000	.000
	3	.991	.917	.549	.202	.046	.006	.000	.000	.000	.000	.000
	4	.999	.978	.758	.389	.126	.025	.003	.000	.000	.000	.000
	5	1.000	.995	.894	.597	.264	.072	.011	.001	.000	.000	.000
	6	1.000	.999	.962	.775	.448	.166	.035	.003	.000	.000	.000
	7	1.000	1.000	.989	.895	.641	.315	.092	.013	.000	.000	.000
	8	1.000	1.000	.997	.960	.801	.500	.199	.040	.003	.000	.000
	9	1.000	1.000	1.000	.987	.908	.685	.359	.105	.011	.000	.000
	10	1.000	1.000	1.000	.997	.965	.834	.552	.225	.038	.001	.000
	11	1.000	1.000	1.000	.999	.989	.928	.736	.403	.106	.005	.000
	12	1.000	1.000	1.000	1.000	.997	.975	.874	.611	.242	.022	.001
	13	1.000	1.000	1.000	1.000	1.000	.994	.954	.798	.451	.083	.009
	14	1.000	1.000	1.000	1.000	1.000	.999	.988	.923	.690	.238	.050
	15	1.000	1.000	1.000	1.000	1.000	1.000	.998	.981	.882	.518	.208
	16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.977	.833	.582
	17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 18$	0	.397	.150	.018	.002	.000	.000	.000	.000	.000	.000	.000
	1	.774	.450	.099	.014	.001	.000	.000	.000	.000	.000	.000
	2	.942	.734	.271	.060	.008	.001	.000	.000	.000	.000	.000
	3	.989	.902	.501	.165	.033	.004	.000	.000	.000	.000	.000
	4	.998	.972	.716	.333	.094	.015	.001	.000	.000	.000	.000
	5	1.000	.994	.867	.534	.209	.048	.006	.000	.000	.000	.000
	6	1.000	.999	.949	.722	.374	.119	.020	.001	.000	.000	.000
	7	1.000	1.000	.984	.859	.563	.240	.058	.006	.000	.000	.000
	8	1.000	1.000	.996	.940	.737	.407	.135	.021	.001	.000	.000
	9	1.000	1.000	.999	.979	.865	.593	.263	.060	.004	.000	.000
	10	1.000	1.000	1.000	.994	.942	.760	.437	.141	.016	.000	.000
	11	1.000	1.000	1.000	.999	.980	.881	.626	.278	.051	.001	.000
	12	1.000	1.000	1.000	1.000	.994	.952	.791	.466	.133	.006	.000
	13	1.000	1.000	1.000	1.000	.999	.985	.906	.667	.284	.028	.002
	14	1.000	1.000	1.000	1.000	1.000	.996	.967	.835	.499	.098	.011

TABLA E

(continuación)

		p									
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	k											
	15	1.000	1.000	1.000	1.000	1.000	.999	.992	.940	.729	.266	.058
	16	1.000	1.000	1.000	1.000	1.000	1.000	.999	.986	.901	.550	.226
	17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.982	.850	.603
	18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=19$	0	.377	.135	.014	.001	.000	.000	.000	.000	.000	.000	.000
	1	.755	.420	.083	.010	.001	.000	.000	.000	.000	.000	.000
	2	.933	.705	.237	.046	.005	.000	.000	.000	.000	.000	.000
	3	.987	.885	.455	.133	.023	.002	.000	.000	.000	.000	.000
	4	.998	.965	.673	.282	.070	.010	.001	.000	.000	.000	.000
	5	1.000	.991	.837	.474	.163	.032	.003	.000	.000	.000	.000
	6	1.000	.998	.932	.666	.308	.084	.012	.001	.000	.000	.000
	7	1.000	1.000	.977	.818	.488	.180	.035	.003	.000	.000	.000
	8	1.000	1.000	.993	.916	.667	.324	.088	.011	.000	.000	.000
	9	1.000	1.000	.998	.967	.814	.500	.186	.033	.002	.000	.000
	10	1.000	1.000	1.000	.989	.912	.676	.333	.084	.007	.000	.000
	11	1.000	1.000	1.000	.997	.965	.820	.512	.182	.023	.000	.000
	12	1.000	1.000	1.000	.999	.988	.916	.692	.334	.068	.002	.000
	13	1.000	1.000	1.000	1.000	.997	.968	.837	.526	.163	.009	.000
	14	1.000	1.000	1.000	1.000	.999	.990	.930	.718	.327	.035	.002
	15	1.000	1.000	1.000	1.000	1.000	.998	.977	.867	.545	.115	.013
	16	1.000	1.000	1.000	1.000	1.000	1.000	.995	.954	.763	.295	.067
	17	1.000	1.000	1.000	1.000	1.000	1.000	.999	.990	.917	.580	.245
	18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.986	.865	.623
	19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=20$	0	.358	.122	.012	.001	.000	.000	.000	.000	.000	.000	.000
	1	.736	.392	.069	.008	.001	.000	.000	.000	.000	.000	.000
	2	.925	.677	.206	.035	.004	.000	.000	.000	.000	.000	.000
	3	.984	.867	.411	.107	.016	.001	.000	.000	.000	.000	.000
	4	.997	.957	.630	.238	.051	.006	.000	.000	.000	.000	.000
	5	1.000	.989	.804	.416	.126	.021	.002	.000	.000	.000	.000
	6	1.000	.998	.913	.608	.250	.058	.006	.000	.000	.000	.000
	7	1.000	1.000	.968	.772	.416	.132	.021	.001	.000	.000	.000
	8	1.000	1.000	.990	.887	.596	.252	.057	.005	.000	.000	.000
	9	1.000	1.000	.997	.952	.755	.412	.128	.017	.001	.000	.000
	10	1.000	1.000	.999	.983	.872	.588	.245	.048	.003	.000	.000
	11	1.000	1.000	1.000	.995	.943	.748	.404	.113	.010	.000	.000
	12	1.000	1.000	1.000	.999	.979	.868	.584	.228	.032	.000	.000
	13	1.000	1.000	1.000	1.000	.994	.942	.750	.392	.087	.002	.000
	14	1.000	1.000	1.000	1.000	.998	.979	.874	.584	.196	.011	.000

TABLA E

(continuación)

		p									
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	k											
	15	1.000	1.000	1.000	1.000	1.000	.994	.949	.762	.370	.043	.003
	16	1.000	1.000	1.000	1.000	1.000	.999	.984	.893	.589	.133	.016
	17	1.000	1.000	1.000	1.000	1.000	1.000	.996	.965	.794	.323	.075
	18	1.000	1.000	1.000	1.000	1.000	1.000	.999	.992	.931	.608	.264
	19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.988	.878	.642
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=25$	0	.277	.072	.004	.000	.000	.000	.000	.000	.000	.000	.000
	1	.642	.271	.027	.002	.000	.000	.000	.000	.000	.000	.000
	2	.873	.537	.098	.009	.000	.000	.000	.000	.000	.000	.000
	3	.966	.764	.234	.033	.002	.000	.000	.000	.000	.000	.000
	4	.993	.902	.421	.090	.009	.000	.000	.000	.000	.000	.000
	5	.999	.967	.617	.193	.029	.002	.000	.000	.000	.000	.000
	6	1.000	.991	.780	.341	.074	.007	.000	.000	.000	.000	.000
	7	1.000	.998	.891	.512	.154	.022	.001	.000	.000	.000	.000
	8	1.000	1.000	.953	.677	.274	.054	.004	.000	.000	.000	.000
	9	1.000	1.000	.983	.811	.425	.115	.013	.000	.000	.000	.000
	10	1.000	1.000	.994	.902	.586	.212	.034	.002	.000	.000	.000
	11	1.000	1.000	.998	.956	.732	.345	.078	.006	.000	.000	.000
	12	1.000	1.000	1.000	.983	.846	.500	.154	.017	.000	.000	.000
	13	1.000	1.000	1.000	.994	.922	.655	.268	.044	.002	.000	.000
	14	1.000	1.000	1.000	.998	.966	.788	.414	.098	.006	.000	.000
	15	1.000	1.000	1.000	1.000	.987	.885	.575	.189	.017	.000	.000
	16	1.000	1.000	1.000	1.000	.996	.946	.726	.323	.047	.000	.000
	17	1.000	1.000	1.000	1.000	.999	.978	.846	.488	.109	.002	.000
	18	1.000	1.000	1.000	1.000	1.000	.993	.926	.659	.220	.009	.000
	19	1.000	1.000	1.000	1.000	1.000	.998	.971	.807	.383	.033	.001
	20	1.000	1.000	1.000	1.000	1.000	1.000	.991	.910	.579	.098	.007
	21	1.000	1.000	1.000	1.000	1.000	1.000	.998	.967	.766	.236	.034
	22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.902	.463	.127
	23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.973	.729	.358
	24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.996	.928	.723
	25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

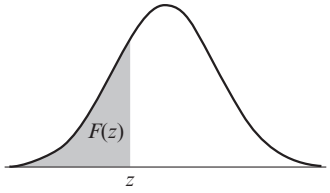
TABLA F

Distribución de dos variables normales con una correlación igual a KR_{21} (tomada de Subkoviak, 1984, pág. 277)

z	KR_{21} = .10	.20	.30	.40	.50	.60	.70	.80	.90
−2.00	.0009	.0014	.0020	.0029	.0041	.0055	.0074	.0098	.0134
−1.90	.0013	.0020	.0030	.0041	.0056	.0075	.0098	.0129	.0173
−1.80	.0020	.0030	.0042	.0058	.0077	.0100	.0130	.0168	.0221
−1.70	.0030	.0043	.0059	.0079	.0104	.0133	.0170	.0216	.0281
−1.60	.0044	.0061	.0083	.0108	.0139	.0175	.0220	.0276	.0353
−1.50	.0063	.0086	.0113	.0145	.0183	.0228	.0282	.0349	.0440
−1.40	.0090	.0119	.0154	.0193	.0239	.0293	.0357	.0436	.0543
−1.30	.0126	.0163	.0205	.0254	.0309	.0374	.0449	.0541	.0664
−1.20	.0173	.0219	.0271	.0329	.0396	.0471	.0559	.0665	.0806
−1.10	.0234	.0291	.0353	.0423	.0500	.0587	.0688	.0809	.0969
−1.00	.0313	.0381	.0455	.0536	.0625	.0725	.0840	.0976	.1155
−0.90	.0413	.0492	.0578	.0671	.0773	.0887	.1015	.1167	.1365
−0.80	.0536	.0628	.0726	.0832	.0947	.1073	.1216	.1383	.1600
−0.70	.0685	.0791	.0902	.1020	.1147	.1286	.1442	.1625	.1860
−0.60	.0865	.0983	.1106	.1237	.1376	.1527	.1696	.1893	.2145
−0.50	.1078	.1207	.1342	.1483	.1633	.1796	.1976	.2186	.2453
−0.40	.1324	.1464	.1609	.1760	.1920	.2092	.2282	.2503	.2784
−0.30	.1606	.1755	.1908	.2067	.2235	.2415	.2614	.2843	.3135
−0.20	.1924	.2079	.2239	.2404	.2577	.2763	.2968	.3204	.3504
−0.10	.2276	.2435	.2598	.2767	.2944	.3134	.3343	.3583	.3888
0.00	.2659	.2821	.2985	.3155	.3333	.3524	.3734	.3976	.4282
0.10	.3072	.3232	.3395	.3564	.3741	.3930	.4139	.4379	.4684
0.20	.3509	.3664	.3824	.3989	.4162	.4348	.4553	.4789	.5089
0.30	.3965	.4113	.4266	.4426	.4593	.4773	.4972	.5202	.5493
0.40	.4433	.4573	.4718	.4869	.5028	.5200	.5391	.5612	.5893
0.50	.4907	.5036	.5171	.5312	.5462	.5625	.5805	.6015	.6283
0.60	.5380	.5498	.5621	.5752	.5891	.6042	.6211	.6408	.6660
0.70	.5846	.5951	.6062	.6181	.6308	.6447	.6603	.6786	.7021
0.80	.6298	.6391	.6489	.6595	.6710	.6836	.6979	.7146	.7363
0.90	.6731	.6811	.6897	.6990	.7092	.7205	.7334	.7486	.7684
1.00	.7140	.7208	.7282	.7363	.7452	.7552	.7667	.7803	.7982
1.10	.7521	.7577	.7640	.7709	.7787	.7874	.7975	.8096	.8255
1.20	.7872	.7918	.7970	.8028	.8094	.8169	.8257	.8363	.8504
1.30	.8190	.8227	.8269	.8318	.8373	.8438	.8513	.8605	.8728
1.40	.8475	.8504	.8538	.8578	.8624	.8678	.8742	.8821	.8928
1.50	.8727	.8750	.8777	.8809	.8847	.8892	.8946	.9012	.9103
1.60	.8948	.8965	.8987	.9012	.9043	.9079	.9124	.9180	.9257
1.70	.9139	.9152	.9168	.9188	.9212	.9242	.9279	.9325	.9389
1.80	.9302	.9311	.9324	.9339	.9358	.9382	.9411	.9449	.9503
1.90	.9439	.9446	.9455	.9467	.9482	.9500	.9524	.9555	.9598

TABLA II

Función de distribución normal tipificada



$$F(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{z} e^{-\frac{x^2}{2}} \cdot dx$$

z	*F*(*z*)	*y*	*z*	*F*(*z*)	*y*	*z*	*F*(*z*)	*y*
–3,00	0,0013	0,0044						
–2,99	0,0014	0,0046	–2,64	0,0041	0,0122	–2,29	0,0110	0,0290
–2,98	0,0014	0,0047	–2,63	0,0043	0,0126	–2,28	0,0113	0,0297
–2,97	0,0015	0,0048	–2,62	0,0044	0,0129	–2,27	0,0116	0,0303
–2,96	0,0015	0,0050	–2,61	0,0045	0,0132	–2,26	0,0119	0,0310
–2,95	0,0016	0,0051	–2,60	0,0047	0,0136	–2,25	0,0122	0,0317
–2,94	0,0016	0,0053	–2,59	0,0048	0,0139	–2,24	0,0125	0,0325
–2,93	0,0017	0,0055	–2,58	0,0049	0,0143	–2,23	0,0129	0,0332
–2,92	0,0018	0,0056	–2,57	0,0051	0,0147	–2,22	0,0132	0,0339
–2,91	0,0018	0,0058	–2,56	0,0052	0,0151	–2,21	0,0136	0,0347
–2,90	0,0019	0,0060	–2,55	0,0054	0,0154	–2,20	0,0139	0,0355
–2,89	0,0019	0,0061	–2,54	0,0055	0,0158	–2,19	0,0143	0,0363
–2,88	0,0020	0,0063	–2,53	0,0057	0,0163	–2,18	0,0146	0,0371
–2,87	0,0021	0,0065	–2,52	0,0059	0,0167	–2,17	0,0150	0,0379
–2,86	0,0021	0,0067	–2,51	0,0060	0,0171	–2,16	0,0154	0,0387
–2,85	0,0022	0,0069	–2,50	0,0062	0,0175	–2,15	0,0158	0,0396
–2,84	0,0023	0,0071	–2,49	0,0064	0,0180	–2,14	0,0162	0,0404
–2,83	0,0023	0,0073	–2,48	0,0066	0,0184	–2,13	0,0166	0,0413
–2,82	0,0024	0,0075	–2,47	0,0068	0,0189	–2,12	0,0170	0,0422
–2,81	0,0025	0,0077	–2,46	0,0069	0,0194	–2,11	0,0174	0,0431
–2,80	0,0026	0,0079	–2,45	0,0071	0,0198	–2,10	0,0179	0,0440
–2,79	0,0026	0,0081	–2,44	0,0073	0,0203	–2,09	0,0183	0,0449
–2,78	0,0027	0,0084	–2,43	0,0075	0,0208	–2,08	0,0188	0,0459
–2,77	0,0028	0,0086	–2,42	0,0078	0,0213	–2,07	0,0192	0,0468
–2,76	0,0029	0,0088	–2,41	0,0080	0,0219	–2,06	0,0197	0,0478
–2,75	0,0030	0,0091	–2,40	0,0082	0,0224	–2,05	0,0202	0,0488
–2,74	0,0031	0,0093	–2,39	0,0084	0,0229	–2,04	0,0207	0,0498
–2,73	0,0032	0,0096	–2,38	0,0087	0,0235	–2,03	0,0212	0,0508
–2,72	0,0033	0,0099	–2,37	0,0089	0,0241	–2,02	0,0217	0,0519
–2,71	0,0034	0,0101	–2,36	0,0091	0,0246	–2,01	0,0222	0,0529
–2,70	0,0035	0,0104	–2,35	0,0094	0,0252	–2,00	0,0228	0,0540
–2,69	0,0036	0,0107	–2,34	0,0096	0,0258	–1,99	0,0233	0,0551
–2,68	0,0037	0,0110	–2,33	0,0099	0,0264	–1,98	0,0239	0,0562
–2,67	0,0038	0,0113	–2,32	0,0102	0,0270	–1,97	0,0244	0,0573
–2,66	0,0039	0,0116	–2,31	0,0104	0,0277	–1,96	0,0250	0,0584
–2,65	0,0040	0,0119	–2,30	0,0107	0,0283	–1,95	0,0256	0,0596

TABLA II *(continuación)*

z	*F*(*z*)	*y*	*z*	*F*(*z*)	*y*	*z*	*F*(*z*)	*y*
–1,94	0,0262	0,0608	–1,49	0,0681	0,1315	–1,04	0,1492	0,2323
–1,93	0,0268	0,0620	–1,48	0,0694	0,1334	–1,03	0,1515	0,2347
–1,92	0,0274	0,0632	–1,47	0,0708	0,1354	–1,02	0,1539	0,2371
–1,91	0,0281	0,0644	–1,46	0,0721	0,1374	–1,01	0,1562	0,2396
–1,90	0,0287	0,0656	–1,45	0,0735	0,1394	–1,00	0,1587	0,2420
–1,89	0,0294	0,0669	–1,44	0,0749	0,1415	–0,99	0,1611	0,2444
–1,88	0,0301	0,0681	–1,43	0,0764	0,1435	–0,98	0,1635	0,2468
–1,87	0,0307	0,0694	–1,42	0,0778	0,1456	–0,97	0,1660	0,2492
–1,86	0,0314	0,0707	–1,41	0,0793	0,1476	–0,96	0,1685	0,2516
–1,85	0,0322	0,0721	–1,40	0,0808	0,1497	–0,95	0,1711	0,2541
–1,84	0,0329	0,0734	–1,39	0,0823	0,1518	–0,94	0,1736	0,2565
–1,83	0,0336	0,0748	–1,38	0,0838	0,1539	–0,93	0,1762	0,2589
–1,82	0,0344	0,0761	–1,37	0,0853	0,1561	–0,92	0,1788	0,2613
–1,81	0,0351	0,0775	–1,36	0,0869	0,1582	–0,91	0,1814	0,2637
–1,80	0,0359	0,0790	–1,35	0,0885	0,1604	–0,90	0,1841	0,2661
–1,79	0,0367	0,0804	–1,34	0,0901	0,1626	–0,89	0,1867	0,2685
–1,78	0,0375	0,0818	–1,33	0,0918	0,1647	–0,88	0,1894	0,2709
–1,77	0,0384	0,0833	–1,32	0,0934	0,1669	–0,87	0,1922	0,2732
–1,76	0,0392	0,0848	–1,31	0,0951	0,1691	–0,86	0,1949	0,2756
–1,75	0,0401	0,0863	–1,30	0,0968	0,1714	–0,85	0,1977	0,2780
–1,74	0,0409	0,0878	–1,29	0,0985	0,1736	–0,84	0,2005	0,2803
–1,73	0,0418	0,0893	–1,28	0,1003	0,1758	–0,83	0,2033	0,2827
–1,72	0,0427	0,0909	–1,27	0,1020	0,1781	–0,82	0,2061	0,2850
–1,71	0,0436	0,0925	–1,26	0,1038	0,1804	–0,81	0,2090	0,2874
–1,70	0,0446	0,0940	–1,25	0,1056	0,1826	–0,80	0,2119	0,2897
–1,69	0,0455	0,0957	–1,24	0,1075	0,1849	–0,79	0,2148	0,2920
–1,68	0,0465	0,0973	–1,23	0,1093	0,1872	–0,78	0,2177	0,2943
–1,67	0,0475	0,0989	–1,22	0,1112	0,1895	–0,77	0,2206	0,2966
–1,66	0,0485	0,1006	–1,21	0,1131	0,1919	–0,76	0,2236	0,2989
–1,65	0,0495	0,1023	–1,20	0,1151	0,1942	–0,75	0,2266	0,3011
–1,64	0,0505	0,1040	–1,19	0,1170	0,1965	–0,74	0,2296	0,3034
–1,63	0,0516	0,1057	–1,18	0,1190	0,1989	–0,73	0,2327	0,3056
–1,62	0,0526	0,1074	–1,17	0,1210	0,2012	–0,72	0,2358	0,3079
–1,61	0,0537	0,1092	–1,16	0,1230	0,2036	–0,71	0,2389	0,3101
–1,60	0,0548	0,1109	–1,15	0,1251	0,2059	–0,70	0,2420	0,3123
–1,59	0,0559	0,1127	–1,14	0,1271	0,2083	–0,69	0,2451	0,3144
–1,58	0,0571	0,1145	–1,13	0,1292	0,2107	–0,68	0,2483	0,3166
–1,57	0,0582	0,1163	–1,12	0,1314	0,2131	–0,67	0,2514	0,3187
–1,56	0,0594	0,1182	–1,11	0,1335	0,2155	–0,66	0,2546	0,3209
–1,55	0,0606	0,1200	–1,10	0,1357	0,2179	–0,65	0,2578	0,3230
–1,54	0,0618	0,1219	–1,09	0,1379	0,2203	–0,64	0,2611	0,3251
–1,53	0,0630	0,1238	–1,08	0,1401	0,2227	–0,63	0,2643	0,3271
–1,52	0,0643	0,1257	–1,07	0,1423	0,2251	–0,62	0,2676	0,3292
–1,51	0,0655	0,1276	–1,06	0,1446	0,2275	–0,61	0,2709	0,3312
–1,50	0,0668	0,1295	–1,05	0,1469	0,2299	–0,60	0,2743	0,3332

TABLA II *(continuación)*

z	$F(z)$	y	z	$F(z)$	y	z	$F(z)$	y
–0,59	0,2776	0,3352	–0,14	0,4443	0,3951	0,31	0,6217	0,3802
–0,58	0,2810	0,3372	–0,13	0,4483	0,3956	0,32	0,6255	0,3790
–0,57	0,2843	0,3391	–0,12	0,4522	0,3961	0,33	0,6293	0,3778
–0,56	0,2877	0,3410	–0,11	0,4562	0,3965	0,34	0,6331	0,3765
–0,55	0,2912	0,3429	–0,10	0,4602	0,3970	0,35	0,6368	0,3752
–0,54	0,2946	0,3448	–0,09	0,4641	0,3973	0,36	0,6406	0,3739
–0,53	0,2981	0,3467	–0,08	0,4681	0,3977	0,37	0,6443	0,3725
–0,52	0,3015	0,3485	–0,07	0,4721	0,3980	0,38	0,6480	0,3712
–0,51	0,3050	0,3503	–0,06	0,4761	0,3982	0,39	0,6517	0,3697
–0,50	0,3085	0,3521	–0,05	0,4801	0,3984	0,40	0,6554	0,3683
–0,49	0,3121	0,3538	–0,04	0,4840	0,3986	0,41	0,6591	0,3668
–0,48	0,3156	0,3555	–0,03	0,4880	0,3988	0,42	0,6628	0,3653
–0,47	0,3192	0,3572	–0,02	0,4920	0,3989	0,43	0,6664	0,3637
–0,46	0,3228	0,3589	–0,01	0,4960	0,3989	0,44	0,6700	0,3621
–0,45	0,3264	0,3605	0,00	0,5000	0,3989	0,45	0,6736	0,3605
–0,44	0,3300	0,3621	0,01	0,5040	0,3989	0,46	0,6772	0,3589
–0,43	0,3336	0,3637	0,02	0,5080	0,3989	0,47	0,6808	0,3572
–0,42	0,3372	0,3653	0,03	0,5120	0,3988	0,48	0,6844	0,3555
–0,41	0,3409	0,3668	0,04	0,5160	0,3986	0,49	0,6879	0,3538
–0,40	0,3446	0,3683	0,05	0,5199	0,3984	0,50	0,6915	0,3521
–0,39	0,3483	0,3697	0,06	0,5239	0,3982	0,51	0,6950	0,3503
–0,38	0,3520	0,3712	0,07	0,5279	0,3980	0,52	0,6985	0,3485
–0,37	0,3557	0,3725	0,08	0,5319	0,3977	0,53	0,7019	0,3467
–0,36	0,3594	0,3739	0,09	0,5359	0,3973	0,54	0,7054	0,3448
–0,35	0,3632	0,3752	0,10	0,5398	0,3970	0,55	0,7088	0,3429
–0,34	0,3669	0,3165	0,11	0,5438	0,3965	0,56	0,7123	0,3410
–0,33	0,3707	0,3778	0,12	0,5478	0,3961	0,57	0,7157	0,3391
–0,32	0,3745	0,3790	0,13	0,5517	0,3956	0,58	0,7190	0,3372
–0,31	0,3783	0,3802	0,14	0,5557	0,3951	0,59	0,7224	0,3352
–0,30	0,3821	0,3814	0,15	0,5596	0,3945	0,60	0,7257	0,3332
–0,29	0,3859	0,3825	0,16	0,5636	0,3939	0,61	0,7291	0,3312
–0,28	0,3897	0,3836	0,17	0,5675	0,3932	0,62	0,7324	0,3292
–0,27	0,3936	0,3847	0,18	0,5714	0,3925	0,63	0,7357	0,3271
–0,26	0,3974	0,3857	0,19	0,5753	0,3918	0,64	0,7389	0,3251
–0,25	0,4013	0,3867	0,20	0,5793	0,3910	0,65	0,7422	0,3230
–0,24	0,4052	0,3876	0,21	0,5832	0,3902	0,66	0,7454	0,3209
–0,23	0,4090	0,3885	0,22	0,5871	0,3894	0,67	0,7486	0,3187
–0,22	0,4129	0,3894	0,23	0,5910	0,3885	0,68	0,7517	0,3166
–0,21	0,4168	0,3902	0,24	0,5948	0,3876	0,69	0,7549	0,3144
–0,20	0,4207	0,3910	0,25	0,5987	0,3867	0,70	0,7580	0,3123
–0,19	0,4247	0,3918	0,26	0,6026	0,3857	0,71	0,7611	0,3101
–0,18	0,4286	0,3925	0,27	0,6064	0,3847	0,72	0,7642	0,3079
–0,17	0,4325	0,3932	0,28	0,6103	0,3836	0,73	0,7673	0,3056
–0,16	0,4364	0,3939	0,29	0,6141	0,3825	0,74	0,7704	0,3034
–0,15	0,4404	0,3945	0,30	0,6179	0,3814	0,75	0,7734	0,3011

TABLA II *(continuación)*

z	*F*(*z*)	*y*	*z*	*F*(*z*)	*y*	*z*	*F*(*z*)	*y*
0,76	0,7764	0,2989	1,21	0,8869	0,1919	1,66	0,9515	0,1006
0,77	0,7794	0,2966	1,22	0,8888	0,1895	1,67	0,9525	0,0989
0,78	0,7823	0,2943	1,23	0,8907	0,1872	1,68	0,9535	0,0973
0,79	0,7852	0,2920	1,24	0,8925	0,1849	1,69	0,9545	0,0957
0,80	0,7881	0,2897	1,25	0,8944	0,1826	1,70	0,9554	0,0940
0,81	0,7910	0,2874	1,26	0,8962	0,1804	1,71	0,9564	0,0925
0,82	0,7939	0,2850	1,27	0,8980	0,1781	1,72	0,9573	0,0909
0,83	0,7967	0,2827	1,28	0,8997	0,1758	1,73	0,9582	0,0893
0,84	0,7995	0,2803	1,29	0,9015	0,1736	1,74	0,9591	0,0878
0,85	0,8023	0,2780	1,30	0,9032	0,1714	1,75	0,9599	0,0863
0,86	0,8051	0,2756	1,31	0,9049	0,1691	1,76	0,9608	0,0848
0,87	0,8078	0,2732	1,32	0,9066	0,1669	1,77	0,9616	0,0833
0,88	0,8106	0,2709	1,33	0,9082	0,1647	1,78	0,9625	0,0818
0,89	0,8133	0,2685	1,34	0,9099	0,1626	1,79	0,9633	0,0804
0,90	0,8159	0,2661	1,35	0,9115	0,1604	1,80	0,9641	0,0790
0,91	0,8186	0,2637	1,36	0,9131	0,1582	1,81	0,9649	0,0775
0,92	0,8212	0,2613	1,37	0,9147	0,1561	1,82	0,9656	0,0761
0,93	0,8238	0,2589	1,38	0,9162	0,1539	1,83	0,9664	0,0748
0,94	0,8264	0,2565	1,39	0,9177	0,1518	1,84	0,9671	0,0734
0,95	0,8289	0,2541	1,40	0,9192	0,1497	1,85	0,9678	0,0721
0,96	0,8315	0,2516	1,41	0,9207	0,1476	1,86	0,9686	0,0707
0,97	0,8340	0,2492	1,42	0,9222	0,1456	1,87	0,9693	0,0694
0,98	0,8365	0,2468	1,43	0,9236	0,1435	1,88	0,9699	0,0681
0,99	0,8389	0,2444	1,44	0,9251	0,1415	1,89	0,9706	0,0669
1,00	0,8413	0,2420	1,45	0,9265	0,1394	1,90	0,9713	0,0656
1,01	0,8438	0,2396	1,46	0,9279	0,1374	1,91	0,9719	0,0644
1,02	0,8461	0,2371	1,47	0,9292	0,1354	1,92	0,9726	0,0632
1,03	0,8485	0,2347	1,48	0,9306	0,1334	1,93	0,9732	0,0620
1,04	0,8508	0,2323	1,49	0,9319	0,1315	1,94	0,9738	0,0608
1,05	0,8531	0,2299	1,50	0,9332	0,1295	1,95	0,9744	0,0596
1,06	0,8554	0,2275	1,51	0,9345	0,1276	1,96	0,9750	0,0584
1,07	0,8577	0,2251	1,52	0,9357	0,1257	1,97	0,9756	0,0573
1,08	0,8599	0,2227	1,53	0,9370	0,1238	1,98	0,9761	0,0562
1,09	0,8621	0,2203	1,54	0,9382	0,1219	1,99	0,9767	0,0551
1,10	0,8643	0,2179	1,55	0,9394	0,1200	2,00	0,9772	0,0540
1,11	0,8665	0,2155	1,56	0,9406	0,1182	2,01	0,9778	0,0529
1,12	0,8686	0,2131	1,57	0,9418	0,1163	2,02	0,9783	0,0519
1,13	0,8708	0,2107	1,58	0,9429	0,1145	2,03	0,9788	0,0508
1,14	0,8729	0,2083	1,59	0,9441	0,1127	2,04	0,9793	0,0498
1,15	0,8749	0,2059	1,60	0,9452	0,1109	2,05	0,9798	0,0488
1,16	0,8770	0,2036	1,61	0,9463	0,1092	2,06	0,9803	0,0478
1,17	0,8790	0,2012	1,62	0,9474	0,1074	2,07	0,9808	0,0468
1,18	0,8810	0,1989	1,63	0,9484	0,1057	2,08	0,9812	0,0459
1,19	0,8830	0,1965	1,64	0,9495	0,1040	2,09	0,9817	0,0449
1,20	0,8849	0,1942	1,65	0,9505	0,1023	2,10	0,9821	0,0440

TABLA II *(continuación)*

z	$F(z)$	y	z	$F(z)$	y	z	$F(z)$	y
2,11	0,9826	0,0431	2,41	0,9920	0,0219	2,71	0,9966	0,0101
2,12	0,9830	0,0422	2,42	0,9922	0,0213	2,72	0,9967	0,0099
2,13	0,9834	0,0413	2,43	0,9925	0,0208	2,73	0,9968	0,0096
2,14	0,9838	0,0404	2,44	0,9927	0,0203	2,74	0,9969	0,0093
2,15	0,9842	0,0396	2,45	0,9929	0,0198	2,75	0,9970	0,0091
2,16	0,9846	0,0387	2,46	0,9931	0,0194	2,76	0,9971	0,0088
2,17	0,9850	0,0379	2,47	0,9932	0,0189	2,77	0,9972	0,0086
2,18	0,9854	0,0371	2,48	0,9934	0,0184	2,78	0,9973	0,0084
2,19	0,9857	0,0363	2,49	0,9936	0,0180	2,79	0,9974	0,0081
2,20	0,9861	0,0355	2,50	0,9938	0,0175	2,80	0,9974	0,0079
2,21	0,9864	0,0347	2,51	0,9940	0,0171	2,81	0,9975	0,0077
2,22	0,9868	0,0339	2,52	0,9941	0,0167	2,82	0,9976	0,0075
2,23	0,9871	0,0332	2,53	0,9943	0,0163	2,83	0,9977	0,0073
2,24	0,9875	0,0325	2,54	0,9945	0,0158	2,84	0,9977	0,0071
2,25	0,9878	0,0317	2,55	0,9946	0,0154	2,85	0,9978	0,0069
2,26	0,9881	0,0310	2,56	0,9948	0,0151	2,86	0,9979	0,0067
2,27	0,9884	0,0303	2,57	0,9949	0,0147	2,87	0,9979	0,0065
2,28	0,9887	0,0297	2,58	0,9951	0,0143	2,88	0,9980	0,0063
2,29	0,9890	0,0290	2,59	0,9952	0,0139	2,89	0,9981	0,0061
2,30	0,9893	0,0283	2,60	0,9953	0,0136	2,90	0,9981	0,0060
2,31	0,9896	0,0277	2,61	0,9955	0,0132	2,91	0,9982	0,0058
2,32	0,9898	0,0270	2,62	0,9956	0,0129	2,92	0,9982	0,0056
2,33	0,9901	0,0264	2,63	0,9957	0,0126	2,93	0,9983	0,0055
2,34	0,9904	0,0258	2,64	0,9959	0,0122	2,94	0,9984	0,0053
2,35	0,9906	0,0252	2,65	0,9960	0,0119	2,95	0,9984	0,0051
2,36	0,9909	0,0246	2,66	0,9961	0,0116	2,96	0,9985	0,0050
2,37	0,9911	0,0241	2,67	0,9962	0,0113	2,97	0,9985	0,0048
2,38	0,9913	0,0235	2,68	0,9963	0,0110	2,98	0,9986	0,0047
2,39	0,9916	0,0229	2,69	0,9964	0,0107	2,99	0,9986	0,0046
2,40	0,9918	0,0224	2,70	0,9965	0,0104	3,00	0,9987	0,0044

Referencias bibliográficas

Abad, F. J., Olea, J., Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud.* Madrid: Síntesis.

Ackerman, T. A. (2005). Multidimensional item response theory modeling. En J. McArdle y A. Maydeu (eds.), *Festschrift for Roc McDonald. Hillsdale.* Nueva York: Erlbaum.

Adams, R. J., Wu, M. L. y Wilson, M. R. (2015). ACER *ConQuest: Generalised Item Response Modelling Software* [computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research. https://www.acer.org/conquest.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40,* 955-959.

Albanese, M. A. (1986). The correction for guessing: A further analysis of Angoff and Schrader. *Journal of Educational Measurement, 23* (3), 225-235.

Allalouf, A. y Shakhar, G. B. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35,* 31-47.

Allen, D. D., Ni, P. y Haley, S. M. (2008). Efficiency and sensitivity of multidimensional computerized adaptive testing of pediatric physical functioning. *Disability and Rehabilitation, 30,* 479-484.

Allen, M. J. y Yen, W. M. (1979). *Introduction to Measurement Theory.* Monterrey, CA: Brooks/Cole Publishing Company.

Alliger, G. M., Lilienfeld, S. O. y Mitchell, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7,* 32-39.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education (1954, 1966, 1974, 1985, 1999, 2014). *Standards for educational and psychological testing.* Washington, DC: APA.

American Federation of Teachers, National Council on Measurement in Education y National Education Association (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: Autor.

American Psychological Association (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Autor.

American Psychological Association (1996). Statement on the disclosure of test data. *American Psychologist, 51,* 644-648.

American Psychological Association (2017). *Ethical principles of psychologists and code of conduct.* Washington, DC: APA.

Amón, J. (1984). *Estadística para psicólogos* (2 vols.). Madrid: Pirámide.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36,* 1086-1093.

Anastasi, A. (1988). *Psychological testing* (2.ª ed.). Nueva York: MacMillan.

Anastasi, A. y Urbina, S. (1997). *Psychological testing* (7.ª ed.). Upper Saddle River, NJ: Prentice-Hall.

Anderberg, M. R. (1973). *Cluster Analysis for Applications.* Nueva York: Academic Press.

Andrich, D. (1988). *Rasch models for measurement.* Beverly Hills, CA: Sage.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. En R. L. Thorndike (ed.), *Educacional mesurement* (2.ª ed.). Washington, DC: American Council on Education.

Angoff, W. H. (1982a). Summary and derivation of equating methods used at ETS. En P. W. Holland y D. R. Rubin (eds.), *Test Equating.* Nueva York: Academic Press.

Angoff, W. H. (1982b). Use of difficulty and discrimination indices for detecting item bias. En R. A. Berk (ed.), *Handbook of methods for detecting test bias.* Baltimore, MD: The Johns Hopkins University Press.

Angoff, W. H. (1984). *Scales, norms and equivalent scores.* Princeton, NJ: Educational Testing Service.

Angoff, W. H. y Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95-105.

Angoff, W. H. y Schrader, W. B. (1984). A study of hypotheses basic to the use of rights and formula scores. *Journal of Educacional Measurement, 21,* 1-17.

Angoff, W. H. y Schrader, W. B. (1986). A rejoinder to Albanese, «The correction for guessing: A further analysis of Angoff and Schrader». *Journal of Educational Measurement, 23* (3), 237-243.

Ansley, T. N. y Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two dimensional data. *Applied Psychological Measurement, 9* (1), 37-48.

Applied Psychological Measurement (1982). Número monográfico sobre TRI, *6* (4), 373-495.

Applied Psychological Measurement (1987). Special series: problems, perspectives and practical issues in equating, *11* (3).

Arendasy, M., Sommer, M., Herle, M., Schützhofer, B. e Inwanschitz, D. (2011). Modeling effects of faking on an objective personality test. *Journal of Individual Differences, 32,* 210-218.

Armayones, M., Boixadós, M., Gómez, B., Guillamón, N., Hernández, E., Nieto, R., Pousada, M. y Sara, B. (2015). Psicología 2.0: Oportunidades y retos para el profesional de la psicología en el ámbito de la e-salud. *Papeles del Psicólogo, 36* (2), 153-160.

Armor, D. J. (1974). Theta reliability and factor scaling. En H. L. Costner (ed.), *Sociological Methodology.* San Francisco, CA: Jossey Bass.

Ascalon, M. E., Meyers, L. S., Davis, B. W. y Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20* (2), 153-170.

Assessment Systems Corporation (1988). *User's manualfor de MicroCAT testing system* (version 3). St. Paul, MN: Autor.

Atkinson, R. C. (ed.) (1964). *Studies in Mathematical Psychology.* Stanford: Stanford University Press.

Ato, M. (1991). *Investigación en ciencias del comportamiento.* Barcelona: PPU.

Ayala, R. J. (2009). *The theory and practice of item response theory.* Nueva York: Guilford Press.

Baker, F. B. (1981). A criticism of Scheuneman's items bias technique. *Journal of Educational Measurement, 18,* 59-62.

Baker, F. B. (1985). *The Basis of Item Response Theory.* Portsmouth, NH: Heineman.

Baker, F. B. (1987). Item parameter estimation under the one, two and three parameter logistic models. *Applied Psychological Measurement, 11* (2), 111-141.

Baldwin, D., Fowles, M. y Livingston, S. (2005). *Guidelines for constructed-response and other performance assessments.* Princeton, NJ: Educational Testing Service.

Barbero, M. (1996). Bancos de ítems. En J. Muñiz (coord.), *Psicometría.* Madrid: Universitas.

Barbero, M. I. (1999). Gestión informatizada de bancos de ítems. En J. Olea, V. Ponsoda y G. Prieto (eds.), *Tests informatizados: fundamentos y aplicaciones.* Madrid: Pirámide.

Barnhart, H. X., Haber, M. y Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics, 58,* 1020-1027.

Barrada, J. R. (2012). Tests adaptativos informatizados: una perspectiva general. *Anales de Psicología, 28* (1), 289-302.

Barth, J. A. (2007). *Automatic processes in social thinking and behavior.* Nueva York: Psychology Press.

Barton, M. A. y Lord, F. M. (1981). An upper asymptote for the three parameter logistic item response model. *Research Bulletin.* Princeton, NJ: Educational Testing Service.

Bartram, D. (1996). Tests qualifications and test use in UK: The competence approach. *European Journal of Psychological Assessment, 12* (1), 62-71.

Bartram, D. (1998). The need for international guidelines on standards for test use: A review of European and international initiatives. *European Psychologist, 2,* 155-163.

Bartram, D. y Coyne, I. (1998). Variation in national patterns of testing and test use: The ITC/EFPPA international survey. *European Journal of Psycological Assessment, 14,* 249-260.

Bartram, D. y Hambleton, R. K. (2006). *Computer-based testing and the Internet: Issues and advances.* Chichester, Inglaterra: Wiley.

Baydar, N. (1990). Effects of coaching on the validity of the SAT: Resu1ts of a simulation study. En W. W. Wilingham, C. Lewis, R. Morgan y L. Ramist (eds.), Predicting colleye grades: An analysis of institutional trends over two decades. Princeton, NJ: Educational Testing Service.

Behavioral and Brain Sciences (1980), *3,* 325-371.

Bell, R. y Lumsden, J. (1980). Test length and validity. *Applied Phychological Measurement, 4* (2), 165-170.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and practice, 18* (3), 5-12.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. En D. Bartram y R. K. Hambleton (eds.), *Computer-based testing and the internet: Issues and advances.* Chichester, Inglaterra: Wiley.

Bergstrom, B. A. y Gershon, R. C. (1995). Item banking. En J. C. Impara (ed.), *Licensure testing: Purposes, procedures, and practices.* Lincoln, NE: Buros.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 15* (4), 4-9.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45* (2), 4-9.

Berk, R. A. (ed.) (1982). *Handbook of methods for detecting test bias.* Baltimore, MD: The Johns Hopkins University Press.

Berk, R. A. (ed.) (1984a). *A guide to criterion-referenced test construction* (2.ª ed.). Baltimore, MD: The Johns Hopkins University Press.

Berk, R. A. (1984b). Selecting the index of reliability. En R. A. Berk (ed.), *A guide to criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research, 56* (1), 137-172.

Berk, R. A. (1996). Standard setting; the next generation (Where few psychometricians have gone before). *Applied Measurement in Education, 9* (3), 21 5-235.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educationaf Measurement, 21,* 147-152.

Binet, A. y Simon, T. H. (1905a). Sur la necessité d'établir un diagnostic scientifique des états inférieurs de l'intelligence. *L'Année Psychologique, 11,* 163-190.

Binet, A. y Simon, T. H. (1905b). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11,* 191-244.

Binet, A. y Simon, T. H. (1908). Le développement de l'intelligence chez les enfants. *L'Année Psychologique, 14,* 1-94.

Birnbaum, A. (1957). Efficient design and use of tests of ability for various decision-making problems. *Series report n.º 58-16. Proyect n.º 7755-23.* USAF School of Aviation Medicine.

Birnbaum, A. (1958a). On the estimation of mental ability. *Series report, n.º 15. Proyect n.º 7755-23.* USAF School of Aviation Medicine.

Birnbaum, A. (1958b). Further considerations of efficiency in tests of a mental ability. *Technical Report n.º 17, Project n.º 7755-23.* Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a examinee's ability. En F. M. Lord y M. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bishop, N. S. y Davis-Becker, S. (2016). Preparing examinees for test taking. En S. Lane, M. R. Raymond y T. M. Haladyna (eds.), *Handbook of test development.* Nueva York: Routledge.

Blanco, M. (1996). *Psicofísica.* Madrid: Universitas.

Bobko, P. (1986). A solution to some dilemmas when testing hypotheses about ordinal interactions. *Journal of Applied Psychology, 71,* 323-326.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16* (4), 21-33.

Bock, R. D. y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443-459.

Bock, R. D. y Wood, R. (1971). Test theory. *Annual Review of Psychology, 22,* 193-224.

Boring, E. G. (1950). *A History of Experimental Psychology.* Nueva York: Appleton (traducción española: México, Trillas, 1978).

Borsboom, D. y Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9,* 91-121.

Borsboom, D. (2005). *Measuring the mind. Conceptual issues in contemporary psychometrics.* Nueva York: Cambridge University Press.

Box, G. E. P. y Draper, N. R. (1987). *Empirical model building and response surfaces.* Nueva York: John Wiley and Sons.

Breithaupt, K. J., Mills, C. N. y Melican, G. J. (2006). Facing the opportunities of the future. En D. Bartram y R. K. Hambleton (eds.), *Computer-based testing and the Internet* (pp. 219-251). Chichester, Inglaterra: John Wiley and Sons.

Brennan, R. L. (1980). Applications of generalizability theory. En R. A. Berk (ed.), *A guide to criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Brennan, R. L. (1983). *Elements of Generalizability Theory.* Iowa City, IA: American College Testing Program.

Brennan, R. L. (1987). Introduction to problems, perspectives and practical issues in equating. *Applied Psychological Measurement, 11* (3), 221-224.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17,* 5-9.

Brennan, R. L. (2001). *Generalizability theory.* Nueva York: Springer.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20* (4), 6-18.

Brennan, R. L. (ed.) (2006). *Educational measurement.* Westport, CT: Praeger.

Brennan, R. L. y Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14,* 277-289.

Brennan, R. L. y Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses and alternatives. *Educational and Psychological Measurement, 41,* 687-699.

Briesch, A. M., Swaminathan, H., Welsh, M. y Chafouleas, S. M. (2014). Generalizability theory: A practi-

cal guide to study design, implementation, and interpretation. *Journal of School Psychology, 52,* 13-35.

Brown, A. y Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18,* 36-52.

Brown, F. G. (1983). *Principles of Education and Psychology Testing.* Nueva York: Holt, Rinehart and Winston.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (2.ª ed.). Nueva York: Guilford Press.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2.ª ed.). Nueva York: Guilford Press.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology, 37,* 1-21.

Buckendahl, C. W. y Plake, B. S. (2006). Evaluating tests. En S. M. Downing y T. M. Haladyna (eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Budescu, D. V. y Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement, 22* (3), 183-196.

Burt, C. (1941). *The Factors of the Mind.* Nueva York: McMillan.

Burt, C. (1955). The reliability estimated by analysis of variance. *British Journal of Statistical Psychology, 8,* 103-118.

Byrne, B. M., Leong, F. T., Hambleton, R. K., Oakland, T., Van de Vijver, F. J. y Cheung, F. M. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3* (2), 94-105.

Cai, L. (2013). *FlexMIRT: Flexible multilevel multidimensional item analysis and test scoring.* Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D. y Du Toit, S. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling.* Lincolnwood, IL: Scientific Software International.

Calero, D. y Padilla, J. L. (2004). Técnicas psicométricas: los tests. En R. Fernández-Ballesteros (ed.), *Evaluación psicológica: conceptos, métodos y estudio de casos* (pp. 323-355). Madrid: Pirámide.

Camilli, G. (1979). A critique of the chi-square method of assessing ítem bias. *Laboratory of educational research.* University of Colorado: Boulder.

Camilli, G. (2006). Test fairness. En R. L. Brennan (ed.), *Educational measurement.* Westport, CT: American Council on Education.

Camilli, G. y Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Campbell, D. T. y Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carifio, J. y Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences, 3* (3), 106-116.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program* (Research Report ONR87-2). Iowa City, IA: American College Testing.

Carmines, E. G. y Zeller, R. A. (1979). *Reliability and Validity Assessment.* Londres: Sage.

Carretero, H. y Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clinical and Health Psychology, 5,* 521-551.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26,* 4, 347-372.

Carver, R. P. (1970). Special problems in measuring change with psychometric device. En *Evaluative research: Strategies and methods.* Pittsburgh, PA: American Institutes for Research.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15,* 373-380.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P. y Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80* (1), 219-251.

Chen, W. H. y Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22* (3), 265-289.

Chernyshenko, O. S. y Stark, S. (2016). Mobile psychological assessment. En F. Drasgow (ed.), *Technology and Testing.* Nueva York: Routledge.

Chien, T. W., Wu, H. M., Wang, W. C., Castillo, R. V. y Chou, W. (2009). Reduction in patient burdens with graphical computerized adaptive testing on the ADL scale: Tool development and simulation. *Health and Quality of Life Outcomes, 5,* 7-39.

Chopin, B. H. (1976). Recent developments in item banking: A review. En D. N. M. Gruijter y L. J. T. van der Kamp (eds.), *Advances in psychological and educational measurement.* Nueva York: Wiley.

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and practice, 15* (2), 20-31.

Cizek, G. J. (ed.) (2012). *Setting performace standards: Foundations, methods and innovations.* Nueva York: Routledge.

Cizek, G. J. y Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standard on tests.* Thousand Oaks, CA: Sage.

Clark, L. A. y Watson, D. (1995). Constructing Validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309-319.

Clauser, B. E. y Clyman, S. G. (1994). A contrasting-groups approach to standard setting for performance assessments of clinical skills. *Academic Medicine, 69* (10), 42-44.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled de sagreement or partial credit. *Psychological Bulletin, 70,* 213-220.

Cohen, J. y Cohen, P. (1983). *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences.* Hillsdale, NY: LEA.

Colom, B. R. (1995). *Tests, inteligencia y personalidad.* Madrid: Pirámide.

Colom, R. (2002). *En los límites de la inteligencia.* Madrid: Pirámide.

Couper, M. P., Tourangeau, R. y Conrad, F. G. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review, 24* (2), 227-245.

Conger, A. J. (1974). A revised definition for suppresor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement, 34,* 35-46.

Coombs, C. H., Dawes, R. M. y Tversky, A. (1981). *Introducción a la psicología matemática.* Madrid: Alianza (orig. 1970).

Crocker, L. (2006). Preparing examinees for test taking: Guidelines for test developers and test users. En S. M. Downing y T. M. Haladyna (eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Crocker, L. y Algina, J. (1986). *Introduction to classical and modern test theory.* Nueva York: Holt, Rinehart and Winston.

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika, 12,* 1-16.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30,* 1-14.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 33,* 116-127.

Cronbach, L. J. (1987). Statistical test for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin, 102* (3), 414-417.

Cronbach, L. I. y Furby, L. (1970). How we should measure «change» or should we? *Psychological Bulletin, 74,* 68-80.

Cronbach, L. J. y Glesser, G. C. (1965). *Psychological Tests and Personnel Decisions.* Urbana, IL: University of Illinois Press.

Cronbach, L. J. y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cronbach, L. J. y Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speed. *Psychometrika, 16,* 167-188.

Cronbach, L. J., Gleser, G. C., Nanda, H. y Rajaratnarn, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles.* Nueva York: Wiley.

Cronbach, L. J., Rajaratnam, N. y Gleser, G. C. (1963). Theory of Generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, 16,* 2, 137-163.

Cuesta, M. (1996). Unidimensionalidad. En J. Muñiz (ed.), *Psicometría* (pp. 239-291). Madrid: Universitas.

Cureton, E. E. (1951). En E. F. Lindquist (ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.

Davey, T. (2011). *Practical considerations in computer-based testing.* Princeton, NJ: Educational Testing Service.

Dawes, R. M. (1972). *Fundamentals of Attitude Measurement.* Nueva York: Wiley.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology, 18,* 87-123.

De Gruijter, D. N. (1980). *Accounting for uncertainty in performance standards.* Documento ERIC, núm. ED 199 280.

De Gruijter, D. N. (1985). Compromise methods for establishing examination standards. *Journal of Educational Measurement, 22,* 263-269.

Delgado, A. R. y Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psycholoyical Assessment, 14* (3), 197-201.

Deng, N. (2009). *References of non-nommercial software for IRT analyses. Center for Educational Assessment Research Report,* núm. 699. Amherst, MA: University of Massachusetts.

Deng, N. y Hambleton, R. K. (2007). *20 Software packages for assessing test dimensionality.* Amherst, MA: University of Massachusetts.

Deville, C. W. (1996). An empirical link of content and construct equivalence. *Applied Psychological Measurement, 20,* 127-139.

Diamond, J. y Evans, W. (1973). The correction for guessing. *Review of Educacional Research, 43,* 181-191.

Dillman, D. A., Smyth, J. D. y Christian, L. M. (2009). *Internet, mail and mixed-mode surveys: The tailored design method,* Hoboken, NJ: John Wiley & Sons.

Donlon, T. (1978). *An Exploratory Study of the Implications of Test Speededness.* Princeton, NY: Educational Testing Service.

Dorans, N. J. y Cook, L. (eds.) (2016). *Fairness in Educational Assessment and Measurement.* Nueva York: Taylor y Francis.

Dorans, N. J. y Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. En P. W. Holland y H. Wainer (eds.), *Differential item funtioning.* Hillsdale, NJ: LEA.

Douglas, J. y Cohen, A. S. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25,* 234-243.

Downing, S. M. (2006). Selected-response item formats in test development. En S. M. Downing y T. M. Haladyna (eds.), *Handbook of test development.* Mahwah, NJ: Erlbaum.

Downing, S. M. (2006). Twelve steps for effective test development. En S. M. Downing y T. M. Haladyna (eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M. y Haladyna, T. M. (2006). *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Draper, N. R. y Smith, H. (1981). *Applied Regression Analysis* (2.ª ed.). Nueva York: Wiley.

Drasgow, F. (ed.) (2016). *Technology and testing.* Nueva York: Routledge.

Drasgow, F. y Parsons, C. K. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189-199.

Drasgow, F., Luecht, R. M. y Bennett, R. E. (2006). Technology and testing. En R. L. Brennan (ed.), *Educational measurement.* Westport, CT: ACE/Praeger.

Du Bois, P. H. (1970). *A History of Psycholoyical Testing.* Boston: Allyn and Bacon.

Dunbar, S. B. y Ordman, V. L. (2003). Validity: Criterion-related. En R. Fernández Ballesteros (ed.), *Encyclopedia of Psychological Assessment* (pp. 1078-1082). Londres: Sage Publications.

Dunlap, W. P. y Kemery, E. R. (1987). Failure to detect moderating effects: Is multicollinearity the problem? *Psychological Bulletin, 102* (3), 418-420.

Dunnete, D. y Borman, W. C. (1979). Personnel selection and classification systems. *Annual Review of Psychology, 30,* 477-525.

Ebel, R. L. (1951). Writing the test item. En E. F. Lindquist (ed.), *Educational Measurement* (pp. 185-249). Washington, DC: American Council on Education.

Ebel, R. L. (1972). *Essentials of Educacional Measurement* (2.ª ed.). Englewood Cliffs, NJ: Prentice-Hall.

Educational Measurement: Issues and Practice (1987), 6 (2). Número especial dedicado a la estrategia «Golden Rule».

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* Nueva York: Academic Press.

Elosua, P. (2003). Sobre la validez de los tests. *Psicothema, 15,* 315-321.

Elosua, P. (2009). ¿Existe vida más allá de SPSS? Descubre R. *Psicothema, 21* (4), 652-665.

Elosua, P. y Geisinger, K. F. (2016). Cuarta evaluación de tests editados en España: forma y fondo. *Papeles del Psicólogo, 37,* 82-88.

Elosua, P. y López, A. (2002). Indicadores de dimensionalidad para ítems binarios. *Metodología de las Ciencias del Comportamiento, 4,* 121-137.

Elosua, P. y Zumbo, B. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema, 20,* 896-901.

Embretson, S. y Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: LEA.

Erceg-Hurn, D. M. y Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63,* 591-601.

Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology, 12,* 263-282.

European Federation of Professional Psychologists Association (1996). Meta-Code of Ethics. *European Psychologist, 1,* 151-154.

Everitt, B. S. (1974). *Cluster Analysis.* Londres: Halstead Press.

Everitt, B. S. (1977). *The Analysis of Contingency Tables.* Londres: Halstead Press.

Evers, A. (1996). Regulations concerning test qualifications and test use in The Netherlands. *European Journal of Psychological Assessment, 12,* 153-159.

Evers, A., McCormick, C., Hawley, L., Muñiz, J. et al. (2017). Testing practices and attitudes toward tests and testing: An international survey. *International Journal of Testing, 17,* 158-190.

Evers, A., Sijtsma, K., Lucassen, W. y Meijer, R. R. (2010). The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results. *International Journal of Testing, 10,* 295-317.

Faggen, J. (1994). *Settiny standards for constructed response tests: An overview.* Princeton, NJ: Educational Testing Service.

Faulkner-Bond, M. y Wells, C. S. (2016). A brief history of and introduction to item response theory. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement: From foundations to future.* Nueva York: Guilford Press.

Fazio, R. H. y Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54,* 297-327.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357-370.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two test. *Psychometrika, 34,* 363-373.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach Alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45,* 99-105.

Feldt, L. S. y Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33,* 141-156.

Feldt, L. S., Steffan, M. y Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351-361.

Feldt, L. S., Woodruff, D. J. y Salih, F. A. (1987). Statistical inference for coefficíent alpha. *Applied Psychological Measurement, 11* (1), 93-103.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika, 7,* 19-29.

Fernández-Ballesteros, R., De Bruyn, E. E. J., Godoy, A., Hornke, L. F., Ter Laak, J., Vizcarro, C., Westhoff, K., Westmeyer, H. y Zaccagnini, J. (2003). Guías para el proceso de evaluación (GAP): una propuesta a discusión. *Papeles del Psicólogo, 23* (84), 58-70.

Ferrando, P. J. y Anguiano, C. (2010). El análisis factorial como técnica de investigación en psicología. *Papeles del Psicólogo, 31,* 18-33.

Ferrando, P. J. y Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema, 29* (2), 236-240.

Fidalgo, A. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (ed.), *Psicometría.* Madrid: Universitas.

Fidalgo, A. y Muñiz, J. (2002). Investigaciones actuales sobre el funcionamiento diferencial de los ítems. *Metodología de las Ciencias del Comportamiento, 4,* 55-66.

Fienberg, S. (1977). *The Analysis of Cross-classified Categorical Data.* Cambridge, MA: MIT Press.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59,* 315-328.

Flanagan, J. C. (1937). A note on calculating the standard error of measurement and reliability coefficients with the test scoring machine. *Journal of Applied Psychology, 23,* 529.

Fleiss, J. L., Cohen, J. y Everitt, B. S. (1969). Large sample standard errors of Kappa and weighted Kappa. *Psychological Bulletin, 72,* 323-327.

Fonseca, E. (2017). Análisis de redes: ¿una nueva forma de comprender la psicopatología? *Revista de Psiquiatría y Salud Mental.* https://doi.org/10.1016/j.rpsm.2017.06.004.

Fonseca, E., Menéndez, L. F., Paino, M., Lemos, S. y Muñiz, J. (2013). Development of a computerized adaptive test for schizotypy assessment. PLoS ONE 8(9): e73201. Doi: 10.1371.

Fonseca, E. y Muñiz, J. (2017). Quinta evaluación de tests editados en España: mirando hacia atrás, construyendo el futuro. *Papeles del Psicólogo, 38* (3), 161-16.

Foster, D. (2016). Testing technology and its effects on test security. En F. Drasgow (ed.), *Technology and testing.* Nueva York: Routledge.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement, 4* (1), 79-90.

Frary, R. B., Tideman, T. N. y Watts, T. M. (1977). Indices of cheating on multiple choice tests. *Journal of Educacional Statistics,* 2, 235-256.

Fraser, C. y McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269.

Fremer, J. (1996). Promoting high standards for test use: Developments in the United States. *European Journal of Psychological Assessment, 12,* 160-168.

Friedman, H. S. (1983). On shutting one's eyes to face validity. *Psychological Bulletin, 94,* 185-187.

Frisbie, D. A. y Becker, D. F. (1991). An analysis of textbook advice about true-false tests. *Applied Measurement in Education, 4,* 67-83.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87,* 564-567.

Galton, F. (1883). *Inquires into Human Faculty and its Development.* Londres: MacMillan.

García-Cueto, E., Muñiz, J. y Lozano, L. M. (2002). Influencia del número de alternativas en las propiedades psicométricas de los tests. *Metodología de las Ciencias del Comportamiento,* supl., 201-205.

García-Pérez, M. A. (1987). A finite state theory of performance in multiple-choice tests. En E. E. Roskam y R. Suck (eds.), *Progress in Mathematical Psychology, 1.* North Holland: Elsevier Science Publishers.

García-Pérez, M. A. (1989). La corrección del azar en pruebas objetivas: un enfoque basado en una nueva teoría de estados finitos. *Investigaciones Psicológicas, 6,* 33-62.

Gawronsky, B. y Payne, B. K. (2010). *Handbook of implicit social cognition: Measurement, theory, and applications.* Nueva York: Guilford.

Geisinger, K. y Usher-Tate, B. J. (2016). A brief history of educational testing and psychometrics. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement. From foundations to future.* Nueva York: Guilford Press.

Gibbons, J. D., Olkin, I. y Sobel, M. (1979). A subset selection technique for scoring items on a multiple choice test. *Psychometrika, 44,* 259-270.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J. et al. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59* (4), 361-368.

Gierl, M. J. y Haladyna, T. M. (eds.) (2013). *Automatic ítem generation: Theory and practice.* Nueva York: Routledge.

Gierl, M. J., Leighton, J. P. y Tan, X. (2006). Evaluating DETECT classification accuracy and consistency when data display complex structure. *Journal of Educational Measurement, 43* (3), 265-289.

Gierl, M. J. y Haladyna, T. M. (eds.) (2013). *Automatic item generation: Theory and practice.* Nueva York: Routledge.

Gifford, J. A. y Swaminathan, H. (1990). Bias and the effect of priors in bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14* (1), 33-43.

Glas, C. (1990). *RIDA: Rasch incomplete design analysis.* Arnhem: The Netherlands, National Institute for Educational Measurement.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Glaser, R. y Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. En R. Gagné (ed.), *Psychological principies in system development.* Nueva York: Holt, Rinehart and Winston.

Gleser, G. C., Cronbach, L. J. y Rajaratnam (1965). Generability of scores innuenced by multiple sources of variance. *Psychometrika, 30,* 395-418.

Goldstein, G. y Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42,* 139-167.

Gómez, J. (1996). Aportaciones de los modelos de estructuras de covarianza al análisis psicométrico. En J. Muñiz (ed.), *Psicometría* (pp. 456- 554). Madrid: Universitas.

Gómez, J., Hidalgo, M. D. y Gilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo, 31* (1), 75-84.

Goodenough, F. L. (1949). *Mental Testing: Its History, Principles, and Applications.* Nueva York: Rinehart.

Goodman, D. P. y Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of curret practices and suggestions for future research. *Applied Measurement in Education, 17,* 145-220.

Greaud, V. A. (1988). Some effects of applying unidimensional IRT to multidimensional tests. *AERA annual meeting,* Nueva Orleans.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17,* 16-19.

Green, P. E. (1976). *Mathematical Tools for Applied Multivariate Analysis.* Londres: Academic Press.

Green, S. B., Lissitz, R. W. y Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educacional and Psychological Measurement, 37,* 827-838.

Greeno, J. G. (1980). Mathematics in psychology. En P. C. Dodwell (ed.), *New horizons in Psychology* (93-113). Londres: Penguin Books.

Greenwald, A. G. y Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4-27.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. I. y Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17-41.

Grier, J. (1975). The number of alternatives for optimum test reliability. *Journal of Educacional Measurement, 12,* l09-112.

Grier, J. (1976). The optimal number of alternatives at a choice point with travel time considered. *Journal of Mathematical Psychology, 14,* 91-97.

Grosse, M. E. y Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation and the Health Proffesions, 9* (3), 267-285.

Guilford, J. P. (1936, 1954). *Psychometric Methods.* Nueva York: McGraw-Hill.

Guilford, J. P. (1967). *The Nature of Human Intelligence.* Nueva York: McGraw-Hill.

Guion, R. M. y Gibson, W. M. (1988). Personnel selection and placement. *Annual Review of Psychology, 39,* 349-374.

Gulliksen, H. (1950). *Theory of Mental Tests.* Nueva York: Wiley (reimpreso en 1987).

Gustafsson, J. E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model

for dichotomous items. *Educational and Psychological Measurement, 40,* 377-385.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Gwet, K. L. (2014). *Handbook of inter-rater reliability.* Gaithersburg, MD: Advanced Analytics.

Haberman, S. J. (1974). *The Analysis of Frecuency Data.* Chicago: University of Chicago Press.

Haberman, S. J. (1978). *Analysis of Qualitative Data* (2 vols.). Nueva York: Academic Press.

Haberman, S. J., Sinharay, S. y Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika, 78,* 417-440.

Haebara, T. (1980). Equating logistic ability scales by weighted least method. *Japanese Psychological Research, 22,* 144-149.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21,* 16-22.

Hakel, M. D. (1986). Personnel selection and placement. *Annual Review of Psychology, 37,* 351-380.

Hakstian, A. R. y Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219-231.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test item* (3.ª ed.). Hillsdale, NJ: LEA.

Haladyna, T. M., Downing, S. M. y Rodríguez, M. C. (2002). A review of multiple-choice item-writing guidelines. *Applied Measurement in Education, 15* (3), 309-334.

Haladyna, T. M. y Rodríguez, M. C. (2013). *Developing and validating test items.* Nueva York: Routledge.

Haladyna, T. M., Downing, S. M. y Rodríguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15,* 309-333.

Hambleton, R. K. (ed.) (1980). Contributions to criterion-referenced testing technology. *Applied Psychological Measurement, 4* (4), 421-581 (número especial dedicado a los tests referidos al criterio).

Hambleton, R. K. (1980). Test score validity and standard setting methods. En R. A. Berk (ed.), *Criterion-referenced measurement: The state of the art* (pp. 80-123). Baltimore, MD: Johns Hopkins University Press.

Hambleton, R. K. (1983a). *Applications of item respose theory.* Vancouver: Educational Research Institute of British Columbia.

Hambleton, R. K. (1983b). Application of item response models to criterion-referenced assesment. *Applied Psychological Measurement, 7* (1), 33-44.

Hambleton, R. K. (1990). Item Response Theory: Introduction and Bibliography. *Psicothema, 11* (1), 97-107.

Hambleton, R. K. (1994a). The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice, 13* (4), 21-26.

Hambleton, R. K. (2004). Theory, methods and practices in testing for the 21st century. *Psicothema, 16* (4), 696-701.

Hambleton, R. K. (2006). Good practices for identifying differential item functionaing. *Medical Care, 44* (11), 182-188.

Hambleton, R. K. (2006). *Testing practices in the 21st century.* Key Note Address, University of Oviedo, Spain, March 8th.

Hambleton, R. K. (2009). *Predicting future directions in testing practices.* ATP Conference, Palm Springs, February, 22-25, 2009.

Hambleton, R. K. y Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Meeasurement, 10,* 159-170.

Hambleton, R. K. y Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8,* 41-56.

Hambleton, R. K. y Pitoniak, M. J. (2006). Setting performance standards. En R. L. Brennan (ed.), *Educational measurement.* Westport, CT: Praeger.

Hambleton, R. K. y Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2* (4), 313-334.

Hambleton, R. K. y Rovinelli, R. J. (1986). Assessing the dirnensionality of a set of test items. *Applied Psychological Measurement, 10,* 3, 287-302.

Hambleton, R. K. y Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10* (1), 19-38.

Hambleton, R. K. y Swaminathan, H. (1985). *Item Response Theory. Principles and applications.* Boston: Kluwer-Nijhoff.

Hambleton, R. K., Clauser, B. E., Mazor, K. M. y Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9* (1), 1-18.

Hambleton, R. K., Swaminathan, H. y Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hambleton, R. K., Merenda, P. F. y Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Londres: Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., Algina, J. y Coulson, D. (1978). Criterion-referenced testing and measuremente: A review of technical issues and developments. *Review of Educational Research, 48,* 1-47.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31* (5), 457-459.

Han, K.T. y Rudner, L. M. (2016). Decision consistency. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement. From foundations to future.* Nueva York: Guilford Press.

Hanley, J. A. (1987). Standard error of the Kappa statistic. *Psychological Bulletin, 102* (2), 315-321.

Hanson, B. A., Harris, D. J. y Brennan, R. L. (1987). *A comparison of several statistical mechods for examining allegations of copying, ACT Research Report Series 87-15.* Iowa City, IA: American College Testing Program.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11* (2), 91-115.

Hartigan, J. A. (1975). *Clustering Algorithms.* Nueva York: Wiley.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49-78.

Hattie, J. A. (1985). Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hattie, J., Krakowski, K., Roger, J. y Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20,* 1-14.

Hattie, J. A. y Krakowski, K. (1994). DIMENSION: A program to generate unidimensional and multidimensional item data. *Applied Psychological Measurement, 17,* 252.

Heise, D. R. y Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. En E. F. Borgatta y G. W. Bohrnstedt (eds.), *Sociological Methodology*. San Francisco, CA: Jossey Bass.

Hernández, A. y González Romá, V. (2000). Evaluación de matrices multirrasgo-multiocasión a través de modelos factoriales aditivos y multiplicativos. *Psicothema, 12,* 283-287.

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G. y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 37,* 192-197.

Hernández, A., Tomás, I., Ferreres, A. y Lloret, S. (2015). Tercera evaluación de tests editados en España. *Papeles del Psicólogo, 36,* 1-8.

Hidalgo, M. D. y López-Pina, J. A. (2000). Funcionamiento diferencial de los ítems: presente y perspectivas de futuro. *Metodología de las Ciencias del Comportamiento, 2,* 167-182.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics, 32,* 1-49.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H. y Schmitt, M. (2005). A Meta-Analysis on the Correlation Between the Implicit Association Test and Explicit Self-Report Measures. *Personality and Social Psychology Bulletin, 31* (10), 1369-1385.

Hofstee, W. K. (1983). The case for compromise in educational selection and grading. En S. B. Anderson y J. S. Helmick (eds.), *On educacional testing.* San Francisco, CA: Jossey Bass.

Hogan, J., Barrett, P. y Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92* (5), 1270-1285.

Hogan, T. P. y Murphy, G. (2007). Recommendations for preparing and scoring constructed-response ittems: What the experts say. *Applied Measurement in Education, 20* (4), 427-441.

Holland, P. W. (1985). On the study of Differential Item Performance without IRT. *Proceedings of the Military Testing Association,* octubre.

Holland, P. W. y Rubin, D. R. (eds.) (1982). *Test equating.* Nueva York: Academic Press.

Holland, P. W. y Thayer, D. T. (1985). *An alternative definition of the ETS delta seale of item difficulty.* Princeton, NJ: Educational Testing Service, Research Report RR-85-43.

Holland, P. W. y Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure* (pp. 86-99). Princeton, NJ: Educational Testing Service, Research Report.

Holland, P. W. y Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer y H. I. Braun (eds.), *Test validity.* Hillsdale, NJ: LEA.

Holland, P. W. y Wainer, H. (eds.) (1993). *Differential item functioning.* Hillsdale, NJ: LEA.

Horst, P. (1966). *Psychological Measurement and Prediction.* Belmont, CA: Wadsworth.

Hough, L. M. y Oswald, F. L. (2000). Personnel selection: Looking toward the future, remembering the past. *Annual Review of Psychology, 51,* 631-664.

Hoyt, C. (1941). Test relibility obtained by analysis of variance. *Psychometrika, 6,* 153-160.

Hu, L. T. y Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6* (1), 1-55.

Huberty, C. J. (1975). Discriminant analysis. *Review of Educacional Research, 45,* 543-598.

Hulin, C. L., Drasgow, F. y Parsons, C. K. (1983). *Item response theory. Aplication to psychological measurement.* Homewood, IL: Dow Jones-lrvin.

Hutchinson, T. P. (1982). Some theories of performance in multiple-choice test, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology, 35,* 71-89.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educacional Measurement, 13,* 253-264.

Impara, J. C. y Foster, D. (2006). Item and test development strategies to minimize test fraud. En S. M. Downing y T. M. Haladyna (eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

International Test Commission (2017). The ITC Guidelines for Translating and Adapting Test (second edition) [www.IntestCom.Org].

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66* (1), 109-132.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. En R. A. Berk (ed.), *Handbook of methods for detecting test bias.* Baltimore, MD: The Johns Hopkins University Press.

Ironson, G. H. y Subkoviak, M. (1979). A comparation of several methods of assessing item bias. *Journal of Educational Measurement, 16,* 209-225.

Irvine, S. y Kyllonen, P. (eds.) (2002). *Item generation for test development.* Mahwah, NJ: Lawrence Erlbaum.

Izquierdo, I., Olea, J. y Abad, F. J. (2014). Exploratory factor analysis en validation studies: Uses and recommendations. *Psicothema, 26,* 395-400.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competence tests: Theory and application. *Educacional Evaluation and Policy Analysis, 4,* 461-475.

Jaeger, R. M. (1989). Certification of student competence. En R. L. Linn (ed.), *Educational Measurement* (3.ª ed.). Nueva York: MacMillan.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educacional Measurement: Issues and Practice, 10* (2), 3-6.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*, 15-40.

Jennrich, R. I. y Bentler, P. M. (2011). Exploratory bifactor analysis. *Psychometrika, 76* (4), 537-549.

Jensen, A. R. (1969). How much can be boost IQ and scholastic achievment? *Harvard Educational Review, 39,* 1-123.

Jensen, A. R. y Munro, E. (1979). Redaction Time, Movement Time and Intelligence. *Intelligence, 3,* 121-126.

Joint Committeee of Testing Practices (2004). *Code of fair testing practices.* Washington, DC: American Psychological Association.

Joncich, G. (1968). *The Sane Positivist: A Biography of Edward L. Thorndike.* Middletown, CT: Wesleyan University Press.

Jones, R. F. (1986). A comparison of the predictive validity of the MCAT for coached and uncoached students. *Journal of Medical Education, 61,* 335-338.

Jöreskog, K. G. y Sörbom, D. (1976). Statistical models and methods for test-retest situations. En D. N. Gruijter y L. J. Van der Kamp (eds.), *Advances in Psychological and Educacional Measurement.* Nueva York: Wiley.

Juan-Espinosa, M. (1997). *Geografía de la inteligencia humana.* Madrid: Pirámide.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Revíew of Educacional Research, 64,* 425-461.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21,* 31-41.

Kane, M. (2006a). Content-related validity evidence in test development. En S. M. Downing y T. M. Haladyna (eds.), *Handbook of test development.* Mahwah, NJ: LEA.

Kane, M. (2006b). Validation. En R. Brennan (ed.), *Educational measurement.* Westport, CT: Praeger.

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. En S. Lane, M. R. Raymond y T. M. Haladyna (eds.), *Handbook of test development.* Nueva York: Routledge.

Kelley, T. L. (1928). *Crossroads in the Mind of Man.* Stanford, CA: Stanford University Press.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of tests items. *Journal of Educacional Psychology, 30,* 17-24.

Kenny, D. A. (1994). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. En P. E. Shorut y S. T. Fiske (eds.), *Personality research, methods and theory.* Hillsdale, NJ: LEA.

Kerlinger, F. N. y Pedhazur, E. J. (1973). *Multiple Regression in Behavioral Research.* Nueva York: Holt, Rinehart and Winston.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3.ª ed.). Pacific Grove, CA: Brooks/Cole.

Klecka, W. R. (1980). *Discriminant Analysis.* Beverly Hills, CA: Sage.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4.ª ed.). Nueva York: Guilford Press.

Koffier, S. L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educacional Measurement, 17,* 167-178.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement, 7* (4), 29-36.

Kolen, M. J. y Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* Nueva York: Springer.

Koo, T. K y Li, M. Y. (2016). A guideline for selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15* (2), 155-163.

Kopec, J. A., Badii, M., McKenna, M., Lima, V. D., Sayre, E. C. y Dvorak, M. (2008). Computerized adaptive testing in back pain: Validation of the CAT-5DQOL. *Spine, 33,* 1384-1390.

Kosinski, M., Stillwell, D. y Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behaviour. *Proceedings of the National Academy of Sciences (PNAS), 110* (15), 5802-5805.

Krantz, D. H., Atkinson, R. C., Luce, R. D. y Suppes, P. (eds.) (1974). *Contemporary Developmencs in Mathematical Psychology* (vol. 1: Leaming, Memory, and Thinking). San Francisco, CA: Freeman.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28,* 221-238.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537-567.

Krosnick, J. A. y Presser, S. (2010). Question and questionnaire desing. En P. V. Marsden y J. D. Wright (eds.), *Handbook of survey research* (2.ª ed.). Bingley, Inglaterra: Emerald Group.

Kuder, G. F. y Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Laming, D. (1973). *Mathematical Psychology.* Londres: Academic Press.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26,* 127-135.

Lane, S. y Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21,* 23-30.

Lane, S., Parke, C. S. y Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17,* 24-28.

Lane, S., Raymond, M. R. y Haladyna, T. M. (eds.) (2016). *Handbook of test development.* Nueva York: Routledge.

Lasko, T. A., Bhagwat, J. G., Zou, K. H. y Ohno-Machad, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics, 38,* 404-415.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edimburg, 61,* 273-287.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburg, 62,* 74-82.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. En S. A. Stouffer et al., *Measurement and prediction.* Princeton, NJ: Princeton University Press.

Lee, R., Miller, K. J. y Graham, W. K. (1982). Correction for restriction of range and attenuation in criterion related validation studies. *Journal of Applied Psychology, 67* (5), 637-639.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6,* 1-24.

Leew, J. y Mair, P. (2007). An introduction to the special volume on psychometric. *Journal of Statistical Software, 20,* 1-5.

Levy, R., Mislevy, R. y Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33* (7), 519-537.

Liang, T. y Wells, C. S. (2009). A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement, 69,* 913-928.

Liang, T., Wells, C. S. y Hambleton, R. K. (2014). An assessment of the nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement, 28* (2), 115-129.

Liang, T., Han, K. T. y Hambleton, R. K. (2009). ResidPlots-2: Computer software for IRT graphical residual analyses. *Applied Psychological Measurement, 33* (5), 411-412.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psichology, 22,* 1-55.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45,* 255-268.

Linacre, J. M. (2015). *Winsteps Rasch measurement computer program.* Beaverton, OR: Winsteps.com.

Linacre, J. M. y Wright, B. D. (1998). *A user's guide to BIGSTEPS.* http://www.winsteps.com/a/bigsteps.pdf.

Lindquist, E. F. (ed.) (1951). *Educational Measurement.* Washington, DC: American Council on Education.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education.* Boston, MA: Houghton Miffiin.

Linn, R. L. (ed.) (1989). *Educational Measurement.* Nueva York: MacMillan.

Linn, R. L. (ed.) (1989). *Educational Measurement.* Washington, DC: American Council on Education.

Linn, R. L. (1990). Admissions testing: Recommended uses, validity, differential prediction, and coaching. *Applied Measurement in Education, 3,* 297-318.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16,* 14-16.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17,* 28-30.

Linn, R. L. y Harnisch, D. L. (1981). Interaction between item content and groups membership on achievment test items. *Journal of Educational Measurement, 18,* 109-118.

Linn, R. L., Levine, M. V., Hastings, C. N. y Wardrop, J. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159-173.

Lissitz, R. W. (ed.) (2009). *The concept of validity: Revisions, new directions, and applications.* Charlotte, NC: Information Age.

Liu, Y. y Maydeu, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement, 73* (2), 254-274.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9,* 13-26.

Livingston, S. (2009). *Constructed-response test questions: Why we use them, how we score them.* Princeton, NJ: Educational Testing Service.

Livingstone, S. A. y Zieky, M. J. (1982). *Passing scores.* Princeton, NJ: ETS.

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A. y Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. *Anales de Psicología, 30* (3), 1151-1169.

López-Pina, J. A. (1995). *Teoría de respuesta al ítem: fundamentos.* Barcelona: PPU.

López-Pina, J. A. e Hidalgo, M. D. (1996). Bondad de ajuste y teoría de respuesta a los ítems. En J. Muñiz (coord.), *Psicometría.* Madrid: Universitas.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, 7.*

Lord, F. M. (1953a). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-549.

Lord, F. M. (1953b). An application of confidence intervals of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57-75.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

Lord, F. M. (1974). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. *Research Bulletin,* 75-133. Princeton, NJ: ETS.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12,* 7-12.

Lord, F. M. (1977). Optimal number of choices per item, a comparison of four approaches. *Journal of Educacional Measurement, 14* (1), 33-38.

Lord, F. M. (1980). *Applications of item response theory to practicaltesting problems.* Hillsdale, NJ: LEA.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educacional Measurement, 21* (3), 239-243.

Lord, F. M. (1986). Maximum likelihood and bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23* (2), 157-162.

Lord, F. M. y Novick, M. R. (1968). *Statistical theories of mental tests scores.* Reading, MA: Addison-Wesley.

Lord, F. M. y Wingersky, M. S. (1983). Comparison of IRT observed-score and true-score «equating». *Research Bulletin,* 83-86. Princeton, NJ: ETS.

Lorr, M. (1983). *Cluster Analysis for the Social Sciences.* San Francisco, CA: Jossey-Bass.

Lozano, L., García-Cueto, E. y Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4* (2), 73-79.

Lubinski, D. y Humphreys, L. G. (1990). Assessing spurious «Moderator Effects»: Ilustrated substantively with the hypothesized («Synergistic») relation between spatial and mathematical ability. *Psychological Bulletin, 107* (3), 385-393.

Luce, R. D., Bush, R. R. y Galanter, E. (eds.) (1963). *Handbook of Mathematical Psychology.* Nueva York: Wiley.

Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin, 58,* 122-131.

Lunz, M. E. (1997). *Constraints, concerns, alternatives for test disclosure.* Comunicación presentada en el congreso de la NCME, Chicago, marzo.

Magno, C. (2009). Taxonomy of aptitude test items: A guide for item writers. *The International Journal of Educational and Psychological Assessment, 2,* 39-53.

Magnuson, D. (1967). *Test Theory.* Reading, MA: Addison-Wesley (traducción española: México, Trillas, 1972).

Maguire, T., Hattie, J. y Brian, H. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research, 40,* 109-126.

Mantel, N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Marascuilo, L. A. y Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on chi-square statistics. *Journal of Educational Measurement, 18,* 229-248.

Markovetz, A., Blaszkiewicz, K., Montag, C., Switala, C. y Schlaepfer, T. E. (2014). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hipotheses, 82* (4), 405-411.

Markus, K. A. y Borsboom, D. (2013). *Frontiers of test validity theory.* Nueva York: Routledge.

Marsh, H. W. (1988). Multitrait-multimethod analysis. En J. P. Keeves (ed.), *Educational Research, methodology and measurement: An international handbook* (pp. 570-580). Oxford: Pergamon Press.

Marshall, J. L. y Haertel, E. H. (1976). *The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests* (manuscrito no publicado). Milwaukee: University of Wisconsin.

Martínez Arias, M. R., Hernández Lloreda, M. J. y Hernández Loreda, M. V. (2006). *Psicometría.* Madrid: Alianza.

Martínez-Cardeñoso, J. García-Cueto, E. y Muñiz, J. (2000). Efecto del entrenamiento sobre las propiedades psicométricas de los tests. *Psicothema, 12,* 358-362.

Maydeu, A. (1996). Modelos multidimensionales de teoría de respuesta a los ítems. En J. Muñiz (coord.), *Psicometría.* Madrid: Universitas.

Mazor, K. M., Clauser, B. E. y Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443-452.

McClelland, G. H. y Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114* (2), 376-390.

McDonald, R. P. (1970). The theoretical foundations of common factor analysis, principal factor analysis and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23,* 1-21.

McDonald, R. P. (1978). Generalizability in factorable domains: Domain validity and generalizability. *Educational and Psychological Measurement, 38,* 75-79.

McDonald, R. P. (1986). Describing the elephant: Structure and function in multivariate data. *Psychometrika, 51* (4), 513-534.

McGraw, K. O. y Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30-46.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16,* 16-18.

Menéndez, L., Peña, E., Fonseca, E. y Muñiz, J. (2017). Computerized adaptive assessment of organizational climate. *Anales de Psicología, 33* (1), 152-159.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. En H. Wainer y H. I. Braun (eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: LEA.

Messick, S. (1989). Validity. En R. L. Linn (ed.), *Educational Measurement* (3.ª ed.). Nueva York: MacMillan.

Messick, S. y Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin,* 89, 191-216.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100,* 398-407.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7* (3), 221-237.

Milligan, G. W. y Cooper, M. C. (1987). Clustering methods. *Applied Psychological Measurement, 11* (4), 329-354.

Millman, J. y Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement, 21* (4), 315-330.

Millman, J., Bishop, H. y Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25,* 707-726.

Mills, C. N. y Breithaupt, K. J. (2016). Current issues in computer-based testing. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement: From fundations to future.* Nueva York: Guilford Press.

Milis, C. N. y Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. *Applied Measurement in Education, 1,* 261-275.

Mills, C. N., Potenza, M. T., Fremer, J. J. y Ward, W. C. (eds.) (2002). *Computer-based testing: Building the foundation for future assessments.* Hillsdale, NJ: LEA.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51* (2), 177-196.

Mislevy, R. J. y Bock, R. D. (1984). *BIWG: Maximum likelihood item analysis and test scoring with logistic models.* Mooresville, IN: Scientific Software.

Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S. y Most, R. B. (1995). Assessment of test user qualifications. *American Psychologist, 5,* 14-23.

Moreno, R., Martínez, R. J. y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema, 16* (3), 490-497.

Moreno, R., Martínez, R. y Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2,* 65-72.

Moreno, R., Martínez, R. y Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice ítems. *Psicothema, 27,* 388-394.

Morris, J. H., Sherman, J. D. y Mansfield, E. R. (1986). Failures to detect moderating effects with ordinary least squares-moderated multiple regression: some reasons and a remedy. *Psychological Bulletin, 99,* 282-288.

Mosier, C. I. (1940). Psychophysics and mental test theory: I. Fundamental postulates and elementary theorems. *Psychological Review, 47,* 355-366.

Mosier, C. I. (1941). Psychophysics and mental test theory: II. The constant process. *Psychological Review, 48,* 235-249.

Moss, P. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17,* 6-12.

Muckle, T. J. (2016). Web-based item development and banking. En S. Lane, M. R. Raymond y T. M. Haladyna (eds.), *Handbook of test development.* Nueva York: Routledge.

Muñiz, J. (1991). *Introducción a los métodos psicofísicos.* Barcelona: PPU.

Muñiz, J. (coord.) (1996a). *Psicometría.* Madrid: Universitas.

Muñiz, J. (1996b). Reunión en Madrid de la Comisión Europea de Tests. *Papeles del Psicólogo, 65,* 89-91.

Muñiz, J. (1997b). Aspectos éticos y deontológicos de la evaluación psicológica. En A. Cordero (ed.), *Evaluación psicológica en el año 2000* (pp. 307-345). Madrid: TEA Ediciones.

Muñiz, J. (1997a). *Introducción a la teoría de respuesta a los ítems.* Madrid: Pirámide.

Muñiz, J. (1998). La medición de lo psicológico. *Psicothema, 10,* 1-21.

Muñiz, J. (2003). *Teoría clásica de los tests.* Madrid: Pirámide.

Muñiz, J. (2004). La validación de los tests. *Metodología de las ciencias del comportamiento, 5* (2), 121-141.

Muñiz, J. (2012). Perspectivas actuales y retos futuros de la evaluación psicológica. En C. Zúñiga (ed.), *Psicología, sociedad y equidad: aportes y desafíos* (pp. 23-45). Santiago de Chile: Universidad de Chile.

Muñiz, J. (2018). La medición de lo psicológico: perspectivas actuales y retos futuros. En Academia de Psicología de España (ed.), *Psicología para un mundo sostenible,* II. Madrid: Pirámide.

Muñiz, J. y Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12,* 206-219.

Muñiz, J. y Cuesta, M. (1993). El problema de la unidimensionalidad en la medición psicológica. En M. Forns y T. Anguera (eds.), *Aportaciones recientes a la evaluación psicológica* (pp. 51-70). Barcelona: PPU.

Muñiz, J. y Fernández-Hermida, J. R. (2000). La utilización de los tests en España. *Papeles del Psicólogo, 76,* 41-49.

Muñiz, J. y Fernández-Hermida, J. R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests. *Papeles del Psicólogo, 31,* 108-121.

Muñiz, J., Fernández-Hermida, J. R., Fonseca, E., Campillo, A. y Peña, E. (2011). Evaluación de tests editados en España. *Papeles del Psicólogo, 32* (2), 113-128.

Muñiz, J. y Fonseca, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación, 5,* 13-25.

Muñiz, J. y Fonseca, E. (2017). *Construcción de instrumentos de medida en psicología.* Madrid: FOCAD. Consejo General de Psicología de España.

Muñiz, J. y Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de Psicología, 52* (1), 41-66.

Muñiz, J. y Hambleton, R. K. (1996). Directrices para la traducción y adaptación de los tests. *Papeles del psicólogo, 66,* 63-70.

Muñiz, J., Elosua, P. y Hambleton, R. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema, 25,* 151-157.

Muñiz, J., García, A. y Virgós, J. M. (1991). Escala de la Universidad de Oviedo para la evaluación del profesorado. *Psicothema, 3* (2), 269-281.

Muñiz, J., Elosua, P., Padilla, J. L. y Hambleton, R. K. (2016). Test adaptation standards for cross-lingual assessment. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement. From foundations to future* (pp. 291-304). Nueva York: Guilford Press.

Muñiz, J., Fidalgo, A., García-Cueto, E., Martínez, R. y Moreno, R. (2005a). *Análisis de los ítems.* Madrid: La Muralla.

Muñiz, J., García-Cueto, E. y Lozano, L. M. (2005b). Item format and the psychometric properties of the Eysenck Personality Questionnaire. *Personality and Individual Differences, 38* (1), 61-69.

Muñiz, J., Hambleton, R. K. y Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1* (2), 115-135.

Muñiz, J., Hernández, A. y Ponsoda, V. (2015). Nuevas directrices sobre el uso de los tests: investigación, control de calidad y seguridad. *Papeles del Psicólogo, 36* (3), 161-173.

Muñiz, J., Prieto, G., Almeida, L. y Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment, 15,* 151-157.

Muraki, E. y Bock, R. D. (1991). *PARSCALE: Parametric scaling of rating data.* Chicago: Scientific Software International.

Murphy, K. (1993). *Honesty in the workplace.* Pacific Grove, CA: Brooks-Cole.

Murphy, K. R. (ed.) (2003). *Validity generalization: A critical review.* Londres: LEA.

Muthén, B. (1988). Some uses of structural equation modelling to validity studies: Extending IRT to external variables. En H. Wainer y H. I. Braun (eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: LEA.

Navas, M. J. (1996). Equiparación de puntuaciones. En J. Muñiz (ed.), *Psicometría.* Madrid: Universitas.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14* (1), 3-19.

Nelson, B., McGorry, P. D., Wichers, M., Wigman, J. T. W. y Hartmann, J. A. (2017). Moving from static to dynamic models of the onset of mental disorder. *JAMA Psychiatry, 74,* 528-534.

Nering, M. L. y Ostini, R. (eds.) (2010). *Handbook of polytomous item response theory models.* Nueva York: Routledge.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22,* 287-293.

Nitko, A. J. (1984). Defining criterion-referenced tests. En R. A. Berk (ed.), *A guide w criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psycholigy, 3,* 1-18.

Olea, J., Abad, F. y Barrada, J. R. (2010). Tests informatizados y otros nuevos tipos de tests. *Papeles del Psicólogo, 31,* 94-107.

Olea, J. y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (coord.), *Psicometría.* Madrid: Universitas.

Olea, J., Ponsoda, V. y Prieto, G. (eds.) (1999). *Tests informatizados: fundamentos y aplicaciones.* Madrid: Pirámide.

Orlando, M. y Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24* (1), 50-64.

Osterlind, S. J. (1998). *Constructing test items: Multiple-Choice, constructed-response, performance and others formats.* Boston, MA: Kluwer Academic Publishers.

Osterlind, S. J. y Everson, H. T. (2009). *Differential item functioning.* Thousand Oaks, CA: Sage.

Osterlind, S. J. y Merz, W. R. (1994). Building a taxonomy for constructed-response test items. *Educational Assessment, 2* (2), 133-147.

Overall. J. E. y Klett, C. J. (1972). *Applied multivariate analysis.* Nueva York: McGraw-Hill.

Padilla, J. L., Gómez, J., Hidalgo, M. D. y Muñiz, J. (2007). Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los test. *Psicothema, 19,* 173-178.

Padilla, J. L. y Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26,* 136-144.

Parshall, C. G., Spray, J. A., Kalohn, J. C. y Davey, T. (2002). *Practical considerations in computer-based testing.* Nueva York: Springer.

Parshall, C. G., Harmes, J. C., Davey, T. y Pashley, P. (2010). Innovative items for computerized testing. En W. J. van der Linden y C. A. Glas (eds.), *Elemens of adapting testing.* Londres: Springer.

Paz, M. D. (1996). Validez. En J. Muñiz (ed.), *Psicometría* (pp. 49-103). Madrid: Universitas.

Pedhazur, E. J. (1982). *Multiple Regression in Behavioral Research* (2.ª ed.). Nueva York: Holt, Rinehart and Winston.

Peng, C. J. y Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359-368.

Peters (1981). *Basic skills improvement policy implementation guide n.º 3: Standards-setting manual.* Boston, MA: Massachusetts State Department of Education.

Petersen, M. A., Groenvold, M., Aaronson, N., Fayers, P., Sprangers, M. y Bjorner, J. B. (2006). Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments and evaluations. *Quality Life Research, 15,* 315-329.

Phelps, R. (ed.) (2005). *Defending standardized testing.* Londres: LEA.

Phelps, R. (ed.) (2008). *Correcting fallacies about educational and psychological testing.* Washington, DC: APA.

Pitoniak, M. J., Sireci, S. y Luecht, R. M. (2002). A multitrait-multimethod validity investigation of scores from professional licensure exam. *Educational and Psychological Measurement, 62,* 498-516.

Pitoniak, M. J. y Cizek, G. J. (2016). Standard setting. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement. From foundations to future.* Nueva York: Guilford Press.

Plake, B. S., Melican, G. J. y Milis, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice, 10* (2), 15-16.

Ponsoda, V. y Hontangas, P. (2013). Segunda evaluación de tests editados en España. *Papeles del Psicólogo, 34,* 82-90.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5,* 285-301.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16,* 9-13.

Popham, W. J. y Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6,* 1-9.

Powers, D. E. (1985). Effects of coaching on GRE aptitude test scores. *Journal of Educational Measurement, 22,* 121-136.

Powers, D. E. (1986). Relation of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100,* 67-77.

Powers, D. E. (1993). Coaching for the SAT: Summary of the summaries and an update. *Educational Measurement: Issues and Practice, 12,* 24-30.

Prieto, G. y Delgado, A. (1996). Construcción de los ítems. En J. Muñiz (coord.), *Psicometría.* Madrid: Universitas.

Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo, 31,* 67-74.

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77,* 65-71.

Putnam, S. E., Pence, P. y Jaeger, R. M. (1995). A multistage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education, 8* (1), 57-83.

Qualls, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement, 29* (3), 213-225.

R Core Team (2014). *R: A language and environment for statistical computing.* Viena: R Foundation for Statistical Consulting.

Raju, N. S. (1977). A generalization of coeficient alpha. *Psychometrika, 42,* 549-565.

Raju, N. S. (1988). The area between two item characteristic curves. *Psycholmetrika, 53,* 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14* (2), 197-207.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhague: The Danish Institute for Educational Research.

Rauthmann, J. (2011). Not only item content but also item formats is important: Taxonomizing item format approaches. *Social Behavior and Personality, 39* (1), 119-128. Doi: 10.2224/sbp.2011.39.1.119.

Rebollo, P., García-Cueto, E., Zardaín, P. C., Cuervo, J., Martínez, I., Alonso, J., Ferrer, M. y Muñiz, J. (2009). Desarrollo del CAT-Health, primer test adaptativo informatizado para la evaluación de la calidad de vida relacionada con la salud en España. *Medicina Clínica, 133* (7), 241-251.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4* (3), 207-230.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17,* 13-16.

Reckase, M. D. (2009). *Multidimensional item response theory. Statistics for Social and Behavioral Sciences.* Londres: Springer.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educacional Measurement: Issues and Practice, 10* (2), 11-14.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667-696.

Reise, S. P. y Haviland, M. G. (2005). Item Response Theory and the Measurement of Clinical Change. *Journal of Personality Assessment, 84,* 228-238.

Restle, F. y Greeno, J. G. (1970). *Introduction to Mathemalical Psychology.* Reading, MA: Addison-Wesley.

Revuelta, J., Abad, F. y Ponsoda, V. (2006). *Modelos politómicos.* Madrid: La Muralla.

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika, 1,* 33-49.

Ríos, J. y Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26,* 108-116.

Rodríguez, 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.

Rodríguez, M. C. (2016). Selected-response item development. En S. Lane, M. R. Raymond y T. M. Haladyna (eds.), *Handbook of test development.* Nueva York: Routledge.

Rogers, H. J. y Hambleton, R. K. (1989). Evaluation of computer simulated baseline statistics for use in bias studies. *Educational and Psychological Measurement, 49,* 355-369.

Rogers, H. J. y Swaminathan, H. H. (2016). Concepts and methods on research on differential functioning of test items. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement: From foundations to future.* Nueva York: Guilford Press.

Rogers, W. T. y Yang, P. (1996). Test wiseness: Its nature and application. *European Journal of Psychological Assessment, 12,* 247-259.

Roid, G. H. (1984). Generating the test items. En R. A. Berk (ed.), *A guide to criterion-referenced test construction.* Baltimore, MA: The Johns Hopkins University Press.

Roid, G. H. y Haladyna, T. M. (1980). The emergence of an item-writing technology. *Review of Educational Research, 50,* 293-314.

Roid, G. H. y Haladyna, T. M. (1982). *A technology for test item writing.* Nueva York: Academic Press.

Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika, 52* (2), 217-233.

Rowley, G. L. y Traub, R. E. (1977). Formula scoring, number-right scoring, and test taking strategy. *Journal of Educational Measurement, 14,* 15-22.

Rudner, L. M. (1977). An approach to biased item identification using latent trait measurement theory. Reunión anual de la AERA, Nueva York.

Rudner, L. M., Getson, P. R. y Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17* (1), 1-10.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educacional Review, 9,* 99-103.

Rusch, T., Mair, O. y Hatzinger, R. (2016). IRT packages in R. En W. van der Linden (ed.), *Handbook of item response theory.* Boca Ratón, FL: Chamman & Hall/CRC.

Ryan, A. M. y Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65,* 693-717.

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 21,* 7-15.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 17.*

Samejima, F. (1974). Normal ogive model on the continous response level in the multidimensional latent space. *Psychometrika, 39,* 11-121.

San Martín, R. y Pardo, A. (1989). *Psicoestadística.* Madrid: Pirámide.

Saunders, D. R. (1956). Moderator variables in prediction. *Educacional and Psycological Measurement, 16,* 209-222.

Scalise, K. y Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing «intermediate constraint» questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assesment, 4* (6). Retrieved from http://www.jtla.org.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16* (3), 143-152.

Schmeiser, C. B. y Welch, C. (2006). Test development. En R. L. Brennan (ed.), *Educational measurement* (4.ª ed.) (pp. 307-353). Westport, CT: American Council on Education/Praeger.

Schmidt, F. L. y Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psycholoyical Methods, 1* (2), 199-223.

Schmidt, F. L. y Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124* (2), 262-274.

Schmitt, N. y Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10,* 1-22.

Schmittlein, D. C. (1984). Assessing validity and test-retest reliability for «pick K of n» data. *Marketing Science, 3,* 23-40.

Shavelson, R. J., Webb, N. M. y Rowley, G. L. (1989). Generalizability Theory. *American Psychologist, 44* (6), 922-932.

Shavelson, R. J. y Webb, N. (1991). *Generalizability theory.* Beverly Hills, CA: Sage.

Shealy, R. T. y Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. En P. W. Holland y H. Wainer (eds.), *Differential item functioning.* Hillsdale, NJ: LEA.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16,* 5-8.

Shepard, L. A. (1982). Definitions of bias. En R. A. Berk (ed.), *Handbook of methods for detecting test bias.* Baltimore, MD: The Johns Hopkins University Press.

Shepard, L. A., Camilli, G. y Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 67,* 317-375.

Shepard. L. A., Camilli, G. y Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93-128.

Shepard. L. A., Camilli, G. y Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22* (2), 77-105.

Shepard, L., Glaser, R., Linn, R. y Bohrnstedt, G. (1993). *Setting performance standards for achievement tests.* Standford, CA: National Academy of Education.

Shermis, M. D. y Burstein, J. (eds.) (2013). *Handbook of automated essay evaluation. Current applications and new directions.* Nueva York: Routledge.

Shoukri, M. M. (2010). *Measures of interobserver agreement and reliability.* Boca Ratón, FL: Taylor and Francis.

Shrock, S. A. y Coscarelli, W. C. (2007). *Criterion-referenced test development.* San Francisco, CA: Pfeiffer.

Shrout, P. E. y Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Silva, F. (1989). *Evaluación conductual y criterios psicométricos.* Madrid: Pirámide.

Simner, M. L. (1996). Recommendations by the Canadian Psychological Association for improving the North American safeguards that help protect the public against test misure. *European Journal of Psychological Assessment, 12,* 72-82.

Sireci, S. (1998a). The construct of content validity. *Social Indicators Research, 45,* 83-117.

Sireci, S. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5* (4), 299-321.

Sireci, S. G. (2003). Validity: Content. En R. Fernández Ballesteros (ed.), *Encyclopedia of Psychological Assessment* (pp. 1075-1077). Londres: Sage Publications.

Sireci, S. y Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26,* 100-107.

Sireci, S. y Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16,* 17-31.

Sireci, S. y Geisinger, K. F. (1995). Using subject matter experts to assess content representation: A MDS analysis. *Applied Psychological Measurement, 19,* 241-255.

Sireci, S. y Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19,* 170-187.

Sireci, S. y Zenisky, A. L. (2006). Innovative items format in computer-based testing: In pursuit of construct representation. En S. M. Downing y T. M. Haladyna (eds.), *Handbook of test development.* Hillsdale, NJ: LEA.

Sireci, S. y Zenisky, A. L. (2016). Computerized innovative item formats: Achievement and credentialing. En S. Lane, M. R. Raymond y T. M. Haladyna (eds.), *Handbook of test development.* Nueva York: Routledge.

Sireci, S. y Faulkner-Bond, M. (2016). The times they are A-changing, but the song remains the same. Future issues and practices in test validation. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement. From foundations to future* (pp. 435-448). Nueva York: Guilford Press.

Smith, G. T., Fischer, S. y Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment, 15,* 467-477.

Smith, J. E. K. (1976). Analysis of qualitative data. *Annual Review of Psychology, 27,* 487-499.

Smith, S. T. (2005). On construct validity: Issues of method measurement. *Psychological Assessment, 17,* 396-408.

Smits, N., Cuijpers, P. y Van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research, 188,* 147-155.

Spath, H. (1980). *Cluster Analysis Algorithms.* Nueva York: Wiley.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5,* 417-426.

Spearman, C. (1927). *The abilities of man.* Nueva York: McMillan.

Stafford, R. S. (1971). The Speededness Quotient: A new descriptive statistic for test. *Journal of Educational Measurement, 8* (4), 275-277.

Standards for educational and psychological testing (1974, 1985). Washinton, DC: American Psychological Asociation.

Stanley, J. C. (1971). Reliability. En R. L. Thorndike (ed.), *Educational Measurement* (2.ª ed.). Washington, DC: American Council on Education.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 667-680.

Stocking, M. y Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Stone, C. A. y Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40,* 331-352.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52* (4), 589-617.

Stout, W. F., Habing, B., Douglas, J., Kim, H., Roussos, L. y Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331-354.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement, 13,* 265-276.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. En R. A. Berk (ed.), *A guide criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Svetina, D. (2013). Assessing dimensionality of noncompensatory multidimensional item response theory with complex structures. *Educational and Psychological Measurement, 73* (2), 312-338.

Svetina, D. y Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment, 19* (1), 35-57.

Swaminathan, H. (1983). Parameter estimation in item response models. En R. K. Hambleton (ed.), *Applications of item response theory* (pp. 24-44). Vancouver, British Columbia: Educational Research Institute of British Columbia.

Swaminathan, H. y Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175-192.

Swaminathan, H. y Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349-364.

Swaminathan, H. y Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51,* 589-601.

Swaminathan, H., Hambleton, R. K. y Algina. J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263-267.

Swaminathan, H., Hambleton, R. K. y Rogers, J. (2007). Assessing the fit of item response theory models. En C. R. Rao y S. Sinharay (eds.), *Handbook of statistics,* vol. 26 (pp. 683-718). Ámsterdam: North Holland.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.* Mahwah, NJ: LEA.

Taleb, N. N. (2008). *El cisne negro.* Barcelona: Paidós (orig. 2007).

Taleporos, E. (1998). Consequential validity: A pratitioner's perspective. *Educational Measurement: Issues and Practice, 17,* 20-23.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27* (3), 159-203.

Tatsuoka, M. M. (1970). *Discriminant Analysis.* Champaign, IL.: Institute for Personality and Ability Testing.

Taylor, H. C. y Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discusion and tables. *Journal of Applied Psychology, 23,* 565-578.

Tenopyr, M. L. y Oeltjen, P. D. (1982). Personnel Selection and Classification. *Annual Review of Psychology, 33,* 581-618.

Terman, L. M. (1916). *The Measurement of Intelligence.* Boston, MA: Houghton Miffiin.

Thissen, D. M. (1986). *MULTIWG: Item analysis and scoring with multiple category response models* (version 5). Mooresville, IN: Scientific Software.

Thissen, D. y Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501-519.

Thissen, D. y Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51* (4), 567-577.

Thorndike, E. L. (1904). *An Introduction to the Theory of Mental and Social Measurements.* Nueva York: Science Press.

Thorndike, R. L. (1951). Reliability. En E. F. Lindquidt (ed.), *Educational Measurement.* Washington, DC: American Council on Education.

Thorndike, R. L. (ed.) (1971). *Educational Measurement* (2.ª ed.). Washington, DC: American Council on Education.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston, MA: Hougton Mifflin.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16,* 433-451.

Thurstone, L. L. (1927a). A law of comparative judgement. *Psychological Review, 34,* 273-286.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal Social Psychology, 21,* 384-400.

Thurstone, L. L. (1928a). The absolute zero in intelligence measurement. *The Psychological Review, 35,* 175-197.

Thurstone, L. L. (1928b). Attitude can be measured. *American Journal of Sociology, 33,* 529-554.

Thurstone, L. L. (1931). *The Reliability and Validity of Tests.* Ann Arbor, MI: Edward Brothers.

Thurstone, L. L. (1937). Psychology as a quantitative rational science. *Science, 85,* 227-232.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monogaphs.*

Thurstone. L. L. (1947). *Multiple Factor Analysis.* Chicago, IL: University of Chicago Press.

Thurstone, L. L. y Ackerson, L. (1929). The mental growth curve for the Binet tests. *The Journal of Educational Psychology, 20,* 569-583.

Thurstone, L. L. y Chave, E. J. (1929). *The Measurement of Attitudes.* Chicago, IL: University of Chicago Press.

Thurstone, L. L. y Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs, 2.*

Timm, N. H. (1975). *Multivariate Analysis with Applications in Education and Psychology.* Monterrey, CA: Brooks-Cole Publising Co.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. En R. A. Berk (ed.), *Handbook of methods for detecting test bias.* Baltimore, MD: The Johns Hopkins University Press.

Torgerson, W. S. (1958). *Theory and Methods of Scaling.* Nueva York: Wiley.

Townsend, J. T. y Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394-401.

Trabin, T. E. y Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. En D. J. Weiss (ed.), *New horizons in testing.* Nueva York: Academic Press.

Traub, R. E. y Hambleton, R. K. (1972). The effect of scoring instructions and degree of speedness on the validity and reliability of multiple-choice tests. *Educational and Psychological Measurement, 32,* 737-758.

Trull, T. J. y Ebner-Premier, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: Introduction to the special section. *Psychological Assessment, 21,* 457-462.

Trull, T. J. y Ebner-Priemer, U. W. (2013). Ambulatory assessment. *Annual Review of Clinical Psycology, 9,* 151-176.

Tsutakawa, R. K. y Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51* (2), 251-268.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1-13.

Tucker, L. R. (1987). *Developments in classical item analysis methods (ETS Research Report, 87-46).* Princeton, NJ: Educational Testing Service.

Turner, S. P. (1979). The concept of face validity. *Quality and Quantity, 13,* 85-90.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1,* 386-391.

Tzelgov, J. y Stem, I. (1978). Relationship between variables in three variables linear regression and the concept of suppressor. *Educational and Psychological Measurement, 38,* 325-335.

Umar, J. (1999). Item banking. En G. N. Masters y J. P. Keeves (eds.), *Advances in measurement in educational research and assessment.* Nueva York: Pergamon.

Urry, V. W. (1977). Tailored testing: A success application of latent trait theory. *Journal of Educational Measurement, 14,* 181-196.

Vale, C. D. (2006). Computerized item banking. En S. M. Downing y T. M. Haladyna (eds.), *Handbook of test development.* Mahwah, NJ: Erlbaum.

Van Abswoude, A., Van der Ark, L. y Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28* (1), 3-24.

Van der Linden, W. J. y Glas, C. A. (eds.) (2010). *Elements of adaptive testing.* Nueva York: Springer.

Van der Linden, W. J. y Hambleton, R. K. (eds.) (1997). *Handbook of modern item response theory.* Nueva York: Springer-Verlag.

Van der Liden, W. (ed.) (2016). *Handbook of item response theory* (3 volúmenes). Boca Ratón, FL: Chamman & Hall/CRC.

Van Os, J., Delespaul, P., Wigman, J., Myin-Germays, I. y Wichers, M. (2013). Beyond DSM and ICD: Introducing «precision diagnosis» for psychiatry using momentary assessment technology. *World Psychiatry, 12,* 113-117.

Von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking.* Nueva York: Springer.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction? En D. J. Weiss (ed.), *New horizons in testing.* Nueva York: Academic Press.

Wainer, H. (ed.) (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: LEA.

Wainer, H. (2000). CATs: Whither and whence. *Psicológica, 21,* 121-133.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F. y Rose, M. (2007). Development and evaluation of a computer adaptive test for «Anxiety» (Anxiety-CAT). *Quality Life Research, 16* (suppl. 1), 143-155.

Ward, A. W. y Murray, M. (1994). Guidelines for the development of ítem banks. *Educational Measurement: Issus and Practice, 13* (1), 34-39.

Weinberg, S. (2003). *Plantar cara. La ciencia y sus adversarios culturales.* Barcelona: Paidós.

Weiss, D. J. (ed.) (1983). *New horizons in testing.* Nueva York: Academic Press.

Wells, C. S. y Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness of fit in item response theory. *Applied Measurement in Education, 21* (1), 22-40.

Wells, C. S. y Faulkner-Bond, M. (eds.) (2016). Educational measurement. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement: From foundations to future.* Nueva York: Guilford Press.

Wells, C. S., Rios, J. y Faulkner-Bond, M. (2016). Testing assumptions of item response theory models. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement: From foundations to future.* Nueva York: Guilford Press.

Wells, C. S. y Faulkner-Bond, M. (eds.) (2016). *Educational Measurement. From Foundations to Future.* Nueva York: Guilford Press.

Way, W. D. y Robin, F. (2016). The history of computer-based testing. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement: From fundations to future.* Nueva York: Guilford Press.

Whitely, S. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9* (1), 1-26.

Wiggins, G. (1993). Assessment: authenticity, context, and validity. *Phi Delta Kappan, 15,* 200-214.

Wiggins, J. S. (1973). *Personality and Prediction: Principles of Personality Assessment.* Reading, MA: Addison-Wesley.

Wilcox, R. R. (1979). Prediction analysis and the reliability of a mastery test. *Educational and Psychological Measurement, 39,* 825-839.

Wilcox, R. R. (1981). Solving measurement problems with an answer —until— correct procedure. *Applied Psychological Measurement, 5,* 399-414.

Wilcox, R. R. (1982). Some new results on a answer —until— correct procedure. *Journal of Educational Measurement, 19,* 67-74.

Wilcox, R. R. (1983). How do examinees behave when taking multiple-choice tests? *Applied Psychological Measurement, 7* (2), 239-240.

Wilcox, R. R. (1985). Estimating the validity of a multiple-choice test item having K correct alternatives. *Applied Psychological Measurement, 9* (3), 311-316.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Williamson, D. M., Bennett, R. E., Lazer, S., Berstein, J., Foltz, P. W., Landauer, T. K., Rubin, D. P., Way, W.

P. y Sweeney, K. (2010). *Automated scoring for the assessment of common core standards.* Princeton, NJ: Educational Testing Service.

Williamson, D. M., Mislevy, R. J. y Bejar, I. (2006). *Automated scoring of complex tasks in computer-based testing.* Mahwah, NJ: LEA.

Winer, B. J. (1971). *Statistical Principles in Experimental Design.* Nueva York: McGraw-Hill.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. En R. K. Hambleton (ed.), *Applications of item response theory* (pp. 45-56). Vancouver, BC: Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M. A. y Lord, F. M. (1982). *LOGIST 5.0, version 1.0 user's guide.* Princeton, NJ: ETS.

Wissler, C. (1901). Correlation of mental and physical traits. *Psychological Monographs, 3* (16).

Wollack, J. A. y Fremer, J. J. (eds.) (2013). *Handbook of test security.* Nueva York: Routledge.

Wood, R., Wingersky, M. S. y Lord, F. M. (1976). *WGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Research Report 76-6). Princeton, NJ: Educational Testing Service.

Woodruff, D. J. y Feldt, L. S. (1986). Test for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika, 51* (3), 393-413.

Wright, B. D. (1968). *Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service.

Wright, B. D. (1977a). Solving measurement problems with the Rasch model. *Journal of Educacional Measurement, 14,* 97-116.

Wright, B. D. (1977b). Misunderstanding of the Rasch model. *Journal of Educacional Measurement, 14,* 219-226.

Wright, B. D. y Bell, S. R. (1984). Items banks: What, why, how. *Journal of Educacional Measurement, 21* (4), 331-346.

Wright, B. D. y Mead, R. J. (1976). *BICAL Calibrating rating scales with the Rasch model* (Research memorandum, n.º 23). Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D. y Panchapakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B. D. y Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA.

Wright, B. D., Mead, R. J. y Bell, S. R. (1979). *BICAL: A Rasch program for the analysis of dichotomus data.* Chicago, IL: MESA.

Wright, B. D., Mead, R. J. y Draba, R. (1976). *Detecting and correcting item bias with a logistic response model* (Research Mem., n.º 23). Chicago: University Chicago, Dept. Ed.

Yan, D., Von Davier, A. A. y Lewis, C. (2014). *Computerized multistage testing: Theory and applications.* Boca Ratón, FL: CRC Press.

Yela, M. (1987). *Introducción a la teoría de los tests.* Madrid: Facultad de Psicología, Universidad Complutense.

Yela, M. (1990). Evaluar qué y para qué. El problema del criterio. *Papeles del Psicólogo, 46/47,* 50-54.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8* (2), 125-145.

Yen, W. M. y Fitzpatrick, A. R. (2006). Item response theory. En R. L. Brennan (ed.), *Educational measurement* (4.ª ed.). Westport, CT: Praeger.

Younger, M. S. (1979). *Handbook for Linear Regression.* North Scituate, MA: Duxbury Press.

Zachary, R. A. y Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology, 41,* 86-94.

Zedeck, S. (1971). Problems with the use of «moderator» variables. *Psychological Bulletin, 76,* 295-310.

Zedeck, S. y Cascio, W. F. (1984). Psychological Issues in Personnel Decisions. *Annual Review of Psychology, 35,* 461-518.

Zenisky, A. L. y Hambleton, R. K. (2016). A model and good practices for score reporting. En S. Lane, M. R. Raymond y T. M. Haladyna (eds.), *Handbook of test development.* Nueva York: Routledge.

Zenisky, A. L. y Luecht, R. M. (2016). The future of computer-based testing. En C. S. Wells y M. Faulkner-Bond (eds.), *Educational measurement: from fundations to future.* Nueva York: Guilford Press.

Zenisky, A. L., Hambleton, R. K. y Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. En W. J. van der Linden y C. A. Glas (eds.), *Elements of adapting testing.* Nueva York: Springer.

Zhang, J. y Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple sturcture. *Psychometrika, 64* (2), 213-249.

Zhao, Y. y Hambleton, R. K. (2009). Software for IRT analyses: Description and features. *Center for Educational Assessment Research Report, 652.* Amherst, MA: University of Massachusetts.

Zieky, M. J. y Livingston, S. A. (1977). *Manual for setting standards on the basic skills assessment tests.* Princeton, NJ: Educational Testing Service.

Zieky, M. J., Perie, M. y Livingston, S. (2008). *Cutscores: A manual for setting standards of performance on edu-*

cational and occupational tests. Princeton, NJ: Educational Testing Service.

Zumbo, B. D. (2007b). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4* (2), 223-233.

Zumbo, B. D. (2007a). Validity: Foundational issues and statistical methodology. En C. R. Rao y S. Sinharay (eds.), *Handbook of statistics* (vol. 26). Psychometrics (pp. 45-79). Ámsterdam, Holanda: Elsevier Science.

Zumbo, B. y Chan, E. (eds.) (2014). *Validity and validation in social, behavioural and health sciences.* Londres: Springer.

Zumbo, B., Gadermann, A. M. y Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods, 6,* 21-29.

Zwick, W. R. y Velicer, W. F. (1996). Comparison of five rules for determining the number of components to retain. *Pychological Bulletin, 99* (3), 432-442.

TÍTULOS RELACIONADOS

Análisis de datos en psicología I, *J. Botella Ausina, M. Suero Suñe y C. Ximénez Gómez.*

Diseños experimentales en psicología, *M. Ato García y G. Vallejo Seco.*

Estadística para psicólogos I. Estadística descriptiva, *J. Amón Hortelano.*

Estadística para psicólogos II. Probabilidad. Estadística inferencial, *J. Amón Hortelano.*

Fundamentos metodológicos en psicología y ciencias afines, *R. Moreno Rodríguez, R. J. Martínez Cervantes y S. Chacón Moscoso.*

Introducción a los métodos de investigación de la psicología, *A. R. Delgado González y G. Prieto Adánez.*

Métodos de investigación y análisis de datos en ciencias sociales y de la salud, *S. Cubo Delgado, B. Martín Marín y J. L. Ramos Sánchez (coords.).*

Problemas resueltos de análisis de datos, *F. J. Pérez Santamaría, V. Manzano Arrondo y H. Fazeli Khalili.*