Psicometría

Book · December 2013

CITATIONS
0

READS
24,249

7 authors, including:

Julio Meneses
Universitat Oberta de Catalunya
69 PUBLICATIONS **552** CITATIONS
SEE PROFILE

Maite Barrios
University of Barcelona
86 PUBLICATIONS **1,256** CITATIONS
SEE PROFILE

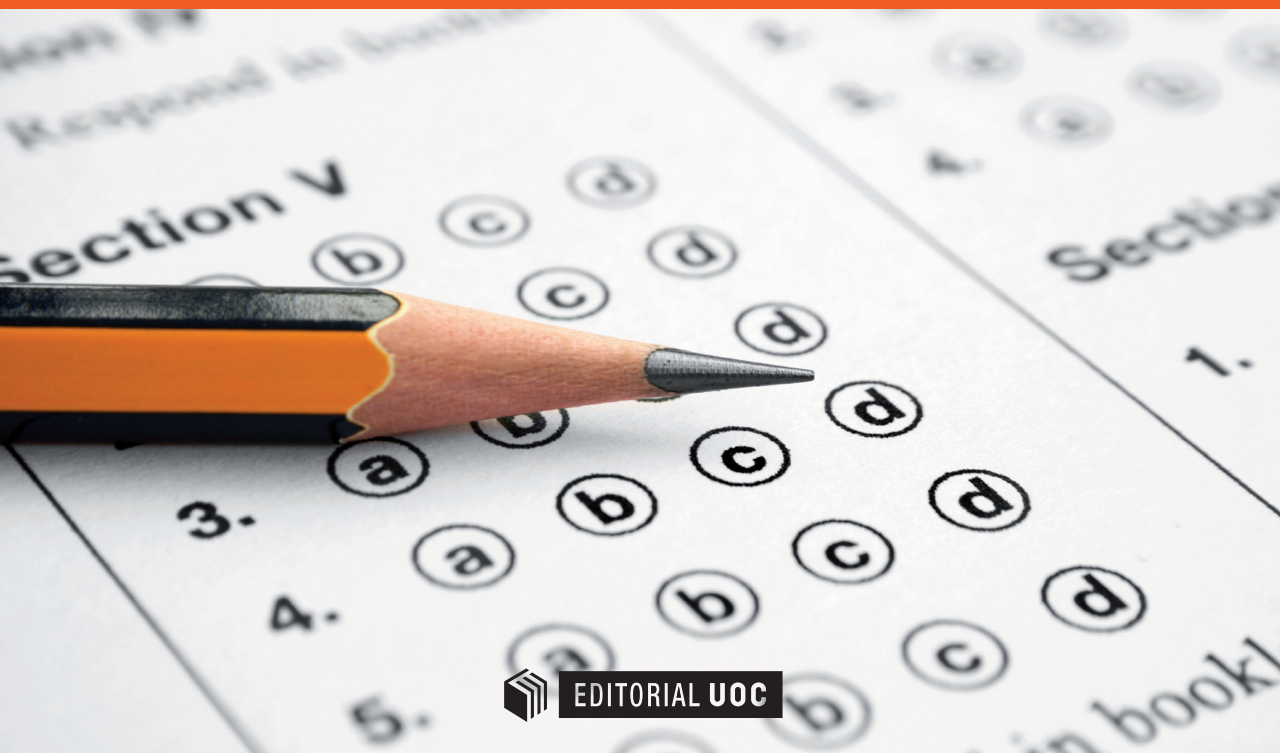
Albert Bonillo
Autonomous University of Barcelona
47 PUBLICATIONS **286** CITATIONS
SEE PROFILE

Luis Manuel Lozano
University of Granada
56 PUBLICATIONS **666** CITATIONS
SEE PROFILE

PSICOLOGÍA

PSICOMETRÍA

JULIO MENESES (COORD.)
MAITE BARRIOS, ALBERT BONILLO,
ANTONI COSCULLUELA, LUIS MANUEL LOZANO,
JAUME TURBANY, SERGI VALERO



EDITORIAL UOC

Psicometría

Psicometría

Julio Meneses (coord.)
Maite Barrios
Albert Bonillo
Antoni Cosculluela
Luis Manuel Lozano
Jaume Turbany
Sergi Valero

EDITORIAL UOC

Diseño de la colección: Editorial UOC

Primera edición en lengua castellana: julio de 2013
Primera edición en formato digital: noviembre de 2013

Gran Via de les Corts Catalanes, 872, 3a Planta - 08018
Barcelona
www.editorialuoc.com

Realización editorial: Eureca Media S.L
ISBN: 978-84-9064-036-4

Julio Meneses

Profesor de los Estudios de Psicología y Ciencias de la Educación de la Universitat Oberta de Catalunya (UOC) e investigador del Internet Interdisciplinary Institute (IN3). Licenciado en Psicología por la Universidad de Oviedo, se graduó en el programa de máster de Sociedad de la información y el conocimiento de la UOC y, actualmente, trabaja en su tesis doctoral. Su investigación se orienta a la aplicación de los métodos de investigación cuantitativa y al uso de técnicas de análisis multivariante, concretamente en las áreas del desarrollo comunitario y la actividad social en la Red, la desigualdad digital, y la relación de los niños y jóvenes con las nuevas tecnologías.

Maite Barrios

Profesora agregada del Departamento de Metodología de las Ciencias del Comportamiento de la Universidad de Barcelona (UB) y consultora de los Estudios de Psicología y Ciencias de la Educación de la Universitat Oberta de Catalunya (UOC). Doctora en Psicología por la UB. Desarrolla su docencia en el ámbito de la metodología impartiendo, entre otras, las asignaturas de Psicometría y Estadística aplicada a la psicología y a la información y documentación. Coautora de varios libros y textos sobre estadística y técnicas de investigación. Su ámbito de trabajo se centra, principalmente, en la informetría, el sesgo de género en ciencia y los estudios bibliométricos.

Albert Bonillo

Profesor lector en el Área de Metodología de las Ciencias del Comportamiento de la Facultad de Psicología de la Universidad Autónoma de Barcelona (UAB) y consultor de los Estudios de Psicología y Ciencias de la Educación de la Universitat Oberta de Catalunya (UOC). Doctor en Psicología por la UAB, máster en Diseño y estadística en ciencias de la salud. En su tesis doctoral propone un conjunto de algoritmos para depurar datos ya registrados. Su investigación se orienta, entre otros ámbitos, a la depuración de datos, la detección de plagio académico, así como a los factores explicativos y las consecuencias de los déficits en la función ejecutiva de los niños.

Antoni Cosculluela

Profesor titular de universidad del Departamento de Metodología de las Ciencias del Comportamiento de la Facultad de Psicología de la Universidad de Barcelona (UB) y consultor de los Estudios de Psicología y Ciencias de la Educación de la Universitat Oberta de Catalunya (UOC). Doctor en Psicología por la UB. Su tarea docente se ha desarrollado en el ámbito de la metodología y los diseños de investigación, y del análisis de datos en psicología y en información y documentación. Autor y coautor de diferentes libros de estadística aplicada. Actualmente su investigación se orienta, entre otros campos, a los estudios bibliométricos en diferentes áreas de conocimiento.

Luis Manuel Lozano

Profesor contratado doctor del Área de Metodología de las Ciencias del Comportamiento de la Universidad de Granada, miembro del CIMCYC y consultor de los Estudios de Psicología de la Universitat Oberta de Catalunya (UOC). Doctor en Psicología por la Universidad de Oviedo. Desarrolla su docencia en el área de metodología, y concretamente imparte las asignaturas de Descripción y exploración de datos en psicología y Técnicas de análisis en investigaciones psicológicas. Sus campos de investigación están relacionados con el análisis frecuentista y bayesiano de datos, la elaboración de cuestionarios y modelos de teoría de la respuesta a los ítems.

Jaume Turbany

Profesor titular de universidad del Departamento de Metodología de las Ciencias del Comportamiento de la Facultad de Psicología de la Universidad de Barcelona (UB) y consultor de los Estudios de Psicología y Ciencias de la Educación de la Universitat Oberta de Catalunya (UOC). Doctor en Psicología por la UB. Su tarea docente se ha desarrollado en el ámbito de la metodología y la estadística aplicadas a la psicología y a la información y documentación. Actualmente su investigación se orienta, entre otros campos, a los estudios bibliométricos en diferentes áreas de conocimiento, y a la investigación en innovación docente.

Sergi Valero

Psicólogo del Servei de Psiquiatria de l'Hospital Universitari Vall d'Hebron de Barcelona y consultor de los Estudios de Psicología y Ciencias de la Educación de la Universitat Oberta de Catalunya (UOC). Doctor en Psicología por la Universidad Autónoma de Barcelona (UAB). Su tarea científica está enmarcada en la investigación básica y aplicada en el ámbito de la psicología, la psicopatología y las demencias. Sus publicaciones incluyen espacios diversos de investigación, desde el estudio delas dimensiones de personalidad normal hasta la genética, pasando por los ensayos clínicos, la construcción y adaptación de instrumentos de medida psicológica, la epidemiología, las drogodependencias o la neuropsicología.

"Mide lo que es medible y haz medible
lo que no lo es".

Atribuido a Galileo Galilei

Índice

Presentación

Julio Meneses

Los tests forman parte de la práctica habitual de los profesionales, docentes e investigadores interesados por la medida indirecta de los fenómenos psicológicos. Al servicio de la evaluación psicológica, y en conjunción con otros instrumentos, como la observación o la entrevista, los tests tienen como propósito principal proporcionar las evidencias necesarias que permitan a los psicólogos y a otros profesionales vinculados con las ciencias sociales y del comportamiento tomar decisiones u orientar sus intervenciones en los diferentes contextos en los que desarrollan su actividad. Los tests cumplen, pues, una importante función en relación con la evaluación y la intervención psicológicas y por ello requieren una atención específica en los programas de formación de los futuros psicólogos. Conocer su funcionamiento, sus propiedades y las condiciones en las que los tests deben ser utilizados de manera adecuada y responsable son algunos de los retos importantes a los que nos enfrentaremos en este texto.

Para hacerlo, a lo largo de los próximos capítulos nos adentraremos en los aspectos teóricos y prácticos involucrados en la medida indirecta de los fenómenos psicológicos mediante tests que la psicometría ha propuesto y sistematizado en las últimas décadas. Como rama de la psicología, la psicometría es la disciplina encargada del desarrollo de teorías, métodos y técnicas que dan apoyo a los procesos de construcción y administración de tests. Su objetivo último, como veremos más adelante, es proporcionar las garantías científicas necesarias para la medida objetiva y estandarizada de los fenómenos psicológicos no observables a partir de una muestra de comportamientos. Este no es un objetivo menor y ha supuesto una importante contribución de la psicología a la teoría de la medida desarrollada en otras disciplinas como la física. Es decir, de acuerdo con las palabras atribuidas a Galileo Galilei en relación con la medida científica: la manera en que la psicometría hace medibles unos fenómenos psicológicos que, por

definición y en oposición a los atributos físicos, no son directamente observables ni manipulables.

Empezaremos este viaje en el capítulo "Aproximación histórica y conceptos básicos de la psicometría", donde situaremos esta disciplina en el contexto general de la psicología, revisando sus antecedentes remotos y más recientes, presentando las aportaciones más relevantes que han contribuido a su establecimiento como disciplina científica y ofreciendo una definición formal que incorpore el papel de los tests en el marco de la evaluación y la intervención psicológicas. A continuación, desarrollaremos los fundamentos de la psicometría, partiendo de una definición y clasificación de los tests, progresando por los diferentes modelos de medida psicométrica desarrollados en las diferentes teorías de los tests, introduciendo brevemente la teoría clásica de los tests y, finalmente, recapitulando estos fundamentos con la revisión del proceso de inferencia psicométrica. Concluiremos el capítulo con una discusión de los procesos de construcción y administración de tests, presentando las principales fases en las que podemos estructurar el diseño y la construcción de un nuevo test, ofreciendo algunos criterios importantes para la evaluación de las características y la valoración de la conveniencia de los tests disponibles en la literatura y, finalmente, abordando los aspectos éticos y deontológicos vinculados al uso de tests en la práctica profesional de la psicología.

En el capítulo "Fiabilidad" trataremos de manera específica un aspecto fundamental para la medida de los fenómenos psicológicos mediante tests: la obtención de puntuaciones consistentes y precisas. Como sucede con cualquier proceso de medida, el desarrollo y la administración de tests requieren el conocimiento del error que se puede cometer. Si este error de medida es grande, de modo que las puntuaciones obtenidas no reflejen adecuadamente los fenómenos psicológicos objeto de interés, los tests no proporcionan la confianza necesaria para cumplir con su objetivo principal al servicio de la evaluación psicológica. Así, en este segundo capítulo abordaremos la fiabilidad a partir de la perspectiva de la teoría clásica de los tests, empezando con una descripción del modelo lineal clásico, derivando el coeficiente de fiabilidad y prestando una especial atención tanto a su interpretación como a las diferentes estrategias que se han ido desarrollando para calcularlo. A continuación, se tratarán tres factores importantes que influyen en la fiabilidad de los tests y se presentarán dos procedimientos para estimar las puntuaciones verdaderas de los sujetos a partir

de las puntuaciones obtenidas. Finalmente, nos ocuparemos de la fiabilidad de los tests referidos a criterio, discutiendo primero sus rasgos principales y presentando, a continuación, los procedimientos clásicos disponibles para evaluar su fiabilidad. Concluiremos el capítulo con una discusión de los métodos más habituales en la determinación de los puntos de corte que permiten la correcta clasificación de los individuos en referencia al criterio.

El capítulo "Validez" abordará otro aspecto clave para la medida indirecta de los fenómenos psicológicos mediante tests: su adecuación a los objetivos para los cuales han sido construidos y son utilizados en la práctica. Partiendo de una revisión histórica, en este capítulo abordaremos las diferentes aproximaciones que la psicometría ha ido proponiendo y definiremos la validez de los tests como el grado en el que la evidencia empírica y la teoría apoyan la interpretación de las puntuaciones en relación con su uso específico. Como veremos a continuación, en la actualidad no se considera la validez como una propiedad intrínseca de los tests, sino que es más bien producto del análisis de su adecuación al propósito específico al que sirven. Para hacerlo, los profesionales interesados por el desarrollo y la administración de tests deben recoger y acumular las evidencias científicas necesarias siguiendo diferentes estrategias. Así, en este capítulo estructuraremos los indicios de validez de los tests en cinco grandes grupos: evidencias basadas en la validez de contenido, basadas en el proceso de respuesta, basadas en la estructura interna del test, basadas en la relación con otras medidas y, finalmente, basadas en las consecuencias de la evaluación. Empezaremos definiendo cada uno de estos indicios de validez y a continuación trataremos en detalle las diferentes estrategias disponibles para obtener las evidencias necesarias. Finalmente, recapitularemos esta discusión tratando los factores que afectan a la validez de los tests teniendo en cuenta su influencia en estas estrategias para obtener los diferentes indicios.

A continuación, en el capítulo "Transformación e interpretación de las puntuaciones" nos ocuparemos de los aspectos metodológicos implicados en el tratamiento de la medida indirecta de los fenómenos psicológicos mediante tests. Más allá de algunas cuestiones teóricas importantes vinculadas al proceso de construcción de tests, en este capítulo discutiremos las operaciones que los profesionales llevan a cabo para hacer interpretables las puntuaciones obtenidas. Como veremos más adelante, estas puntuaciones no son, *per se*, informativas y han de ser siempre interpretadas para hacerlas útiles de acuerdo con el propósi-

to con el que los tests han sido desarrollados. Así, empezaremos presentando el marco general de interpretación de las medidas obtenidas mediante tests y abordaremos algunas estrategias importantes para la transformación de puntuaciones, como son la construcción de percentiles o de puntuaciones estandarizadas y la utilización de normas cronológicas. Estas estrategias, que trataremos en detalle teniendo en cuenta sus características, la manera de calcularlas y sus limitaciones, sirven para recodificar las puntuaciones obtenidas en un nuevo sistema de valores que facilite su interpretación sin afectar a la distinta posición de los sujetos en relación con las magnitudes de las puntuaciones originales. A continuación trataremos el proceso de baremación o escalamiento de la medida, que tiene por objetivo establecer una conexión entre la puntuación del individuo y la ejecución observada en un grupo de referencia. Finalmente, concluiremos el capítulo con una exposición de las diferentes estrategias disponibles para hacer equiparables las puntuaciones que proporcionan tests diferentes que tienen por objetivo la medida de los mismos fenómenos psicológicos.

En el capítulo "Análisis de los ítems" introduciremos brevemente un último aspecto importante para la medida indirecta de los fenómenos psicológicos mediante tests: el análisis del funcionamiento de los ítems que conforman los propios tests. A pesar de que es una cuestión muy relevante para la psicometría, especialmente en el diseño y la construcción de nuevos instrumentos, no siempre forma parte de los programas de formación de los futuros psicólogos. En este capítulo abordaremos el análisis de las propiedades de los ítems en el caso específico de las pruebas de ejecución máxima –también denominadas tests de habilidad o de potencia–, que tienen por objetivo evaluar la competencia, la aptitud o los conocimientos de los individuos a partir del acierto o la calidad de sus respuestas. Como veremos, a diferencia de las pruebas de ejecución típica, este tipo de pruebas discriminan respuestas correctas e incorrectas y es esta la base empleada para puntuar las ejecuciones individuales. Partiendo de una definición de estos dos tipos de pruebas, empezaremos discutiendo algunas directivas importantes para la construcción de los ítems, al tiempo que presentaremos una prueba de ejecución máxima ficticia que nos servirá para ilustrar esta exposición. Así, abordaremos las propiedades de los ítems de acuerdo con la formulación de la teoría clásica de los tests y discutiremos los diferentes procedimientos disponibles para evaluar la dificultad y la discriminación de los ítems. Finalmente, apuntaremos la lógica propuesta por la teoría de la res-

puesta al ítem y concluiremos presentando el desarrollo de los cálculos necesarios para evaluar las propiedades de los ítems que conforman la prueba ficticia que hemos utilizado para ilustrar los diferentes procedimientos.

Esta obra tiene como objetivo general conocer los fundamentos de la psicometría como disciplina científica encargada de la medida indirecta de los fenómenos psicológicos mediante el desarrollo y la administración de tests.

Además del objetivo general, tiene como objetivos específicos:

- Situar la psicometría en el contexto general de la psicología al servicio de la evaluación y la intervención psicológicas.
- Saber definir y clasificar los diferentes tipos de tests disponibles.
- Conocer los diferentes modelos de medida desarrollados por la psicometría en las diferentes teorías de los tests.
- Entender el proceso de inferencia psicométrica y conocer los retos específicos que la medida mediante tests debe afrontar en el contexto del método científico.
- Conocer las implicaciones prácticas de los procesos de construcción y administración de tests, haciendo un énfasis especial en los aspectos éticos y deontológicos vinculados con su uso.
- Entender el concepto de fiabilidad desde la perspectiva psicométrica.
- Saber calcular e interpretar los coeficientes de fiabilidad desde la perspectiva de la teoría clásica de los tests.
- Conocer los factores que afectan a la fiabilidad de una medida.
- Saber estimar las puntuaciones verdaderas de los sujetos a partir de sus puntuaciones empíricas.
- Conocer los procedimientos para abordar la fiabilidad de los tests referidos a criterio.
- Conocer los métodos disponibles para determinar el punto de corte para clasificar a los individuos.
- Conocer los procesos de validación de los tests que permiten inferir su adecuación a los objetivos para los cuales han sido construidos y son utilizados en la práctica.
- Saber definir y clasificar los tipos de validez en función de los diferentes indicios que se pueden recoger como evidencias.

- Conocer de manera práctica las diferentes formas de validez para saber si las conclusiones que se extraen a partir de la aplicación de los tests resultan adecuadas.
- Conocer los factores que afectan a los diferentes tipos de indicios de validez.
- Saber elegir el test más adecuado en función de los indicios disponibles de su validez.
- Desarrollar un punto de vista crítico en la interpretación de las puntuaciones obtenidas mediante tests.
- Conocer las distintas estrategias disponibles para transformar e interpretar las puntuaciones de los tests.
- Conocer las diversas estrategias disponibles para equiparar las puntuaciones obtenidas con diferentes instrumentos que miden los mismos fenómenos psicológicos.
- Entender qué es un baremo y cuáles son los rasgos fundamentales que le otorgan calidad.
- Conocer los diferentes procedimientos disponibles para valorar los ítems de las pruebas de ejecución máxima.
- Conocer las directivas disponibles para la construcción de ítems que conforman las pruebas de ejecución máxima.
- Saber valorar la adecuación de los ítems de las pruebas de ejecución máxima a partir de sus propiedades psicométricas.
- Conocer las diferencias básicas en el análisis de los ítems desde las perspectivas de la teoría clásica de los tests y de la teoría de respuesta al ítem.

Más allá de las referencias citadas en los diferentes capítulos, que sirven para profundizar en algunos aspectos que van más allá de los límites de este texto, a continuación ofrecemos una selección de contribuciones desarrolladas en nuestro contexto inmediato que pueden ser de utilidad para obtener una visión complementaria a esta aproximación a la psicometría.

Fernández-Ballesteros, R. (1997). Evaluación psicológica y tests. En A. Cordero (Ed.), *La evaluación psicológica en el año 2000* (pp. 11-26). Madrid: TEA Ediciones.

El libro editado por TEA en torno a los retos para la evaluación psicológica ofrece algunas lecturas interesantes sobre el papel que desempeñan los tests en el ejercicio profesional de los psicólogos. En este sentido, el capítulo de la profesora Fernández-Ballesteros presenta una excelente síntesis sobre los retos de la medida indirecta de los fenómenos psicológicos mediante tests, abordando una breve aproximación histórica, una discusión sobre los usos terminológicos, el encaje de los tests en el proceso general de la evaluación psicológica y una discusión sobre la validez de las puntuaciones de los tests.

Yela, M. (1996). Los tests. *Psicothema*, 8, supl. 1, 249-263. Disponible en línea en http://www.psicothema.com/pdf/660.pdf.

Este es un texto clásico publicado en el año 1987 por el profesor Yela en su manual *Introducción a la teoría de los tests*, y reproducido en los suplementos de la revista *Psicothema*, donde presenta una definición y descripción general de los tests como reactivos que revelan o dan testimonio fiel de los fenómenos psicológicos no observables. El texto ofrece también una aproximación histórica a su desarrollo, una clasificación de los diferentes tipos de tests disponibles y, finalmente, una discusión sintética de las diferentes fases implicadas en el desarrollo de nuevos instrumentos.

Muñiz, J. (1998). La medición de lo psicológico. *Psicothema*, 10 (1), 1-21. Disponible en línea en http://www.psicothema.com/pdf/138.pdf.

Este texto corresponde con la conferencia inaugural del curso 1997-1998 de la Universidad de Oviedo, donde tuvimos la oportunidad de escuchar la voz autorizada del profesor Muñiz en relación con los importantes retos que supone la medida de los fenómenos psicológicos no observables para la psicología. El artículo empieza con una descripción de las características esenciales de estos fenómenos, aborda los orígenes de la medida mediante tests y discute tres propiedades básicas para un uso adecuado de los tests: la fiabilidad, la validez y la fundamentación teórica. En relación con esta última propiedad, su exposición de los diferentes modelos de medida psicométrica propuestos por Fraser (1980) es una aproximación alternativa interesante a la que hemos desarrollado en nuestro texto.

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, 31 (1), 57-66. Disponible en línea en http://www.papelesdelpsicologo.es/pdf/1796.pdf.

Este número de la revista Papeles del Piscólogo es una importante referencia para los profesionales interesados por la medida indirecta de los fenómenos psicológicos mediante tests. Entre sus artículos, la contribución de Muñiz ofrece una aproximación muy comprensible a las teorías de los tests y el papel que desempeñan en el establecimiento de las inferencias a partir de las puntuaciones obtenidas. Esta exposición arranca con una nota histórica que conduce hacia una caracterización de la teoría clásica de los tests, una discusión de sus limitaciones y, finalmente, presenta las soluciones que la teoría de respuesta al ítem ha propuesto recientemente para hacer frente a estas limitaciones. Más allá de la amena exposición, el lector interesado en conocer estas dos aproximaciones no debería dejar de tener presente la magnífica tabla en la que, en la parte final del artículo, se comparan sus características básicas.

Prieto, G. y Delgado, A. R. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, 31 (1), 67-74. Disponible en línea en http://www.papelesdelpsicologo.es/pdf/1797.pdf.

En el mismo número de la revista, los profesores Prieto y Delgado ofrecen una panorámica de la fiabilidad y la validez de los tests, tanto desde un punto de vista conceptual como atendiendo a los procedimientos más habituales para su evaluación. El artículo discute algunas nociones erróneas sobre estos principios, como son considerar la fiabilidad y la validez propiedades intrínsecas de los tests o considerarlas de manera absoluta y no como una cuestión de grado.

Ferrando, J. y Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación psicológica. *Papeles del Psicólogo*, 31 (1), 18-33. Disponible en línea en http://www.papelesdelpsicologo.es/pdf/1793.pdf.

Una tercera contribución interesante publicada en este número de la revista Papeles del Psicólogo es la de los profesores Ferrando y Anguiano-Carrasco, quienes proponen una aproximación muy accesible al análisis factorial como instrumento de investigación psicológica. Después de una revisión conceptual,

los autores discuten las diferencias principales entre el análisis factorial exploratorio y el análisis factorial confirmatorio, así como presentan los principales procedimientos implicados para estimar los modelos correspondientes. A continuación, ilustran las diferentes etapas implicadas en una investigación, desde su diseño y la recogida de datos hasta la interpretación de la solución final. Se presenta también el programa Factor, un recurso de distribución libre muy interesante para llevar a cabo todos los cálculos implicados en el análisis factorial.

Moreno, R., Martínez, R. J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, 16 (3), 490-497. Disponible en línea en http://www.psicothema.com/pdf/3023.pdf.

Este es un artículo muy interesante para los profesionales interesados por el desarrollo de los ítems de elección múltiple que conforman un nuevo test orientado a la evaluación de competencias, aptitudes o conocimientos. A partir de una revisión crítica de las diferentes directrices existentes, los profesores Moreno, Martínez y Muñiz se proponen unificar la diversidad de criterios y proponen un conjunto de recomendaciones que, bajo el principio de parsimonia, facilite una adecuada redacción de este tipo de ítems. Se presentan un total de doce directrices básicas que los autores, además, ilustran con algunos ejemplos.

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, 77, 65-75. Disponible en línea en http://www.papelesdelpsicologo.es/vernumero.asp?id=1102.

Por su parte, los profesores Prieto y Muñiz presentan en este artículo el modelo de evaluación de los tests desarrollado por el Colegio Oficial de Psicólogos –en la actualidad, Consejo General de Colegios Oficiales de Psicólogos–, donde proponen un procedimiento de evaluación de la calidad y las características de los tests y presentan un cuestionario para sistematizar todo el proceso. Este modelo ha sido ya utilizado con éxito para evaluar diez de los tests más utilizados por los psicólogos españoles y proporciona una guía interesante para conocer las decisiones que los profesionales interesados en la elaboración de tests deben tomar. Este modelo es también una importante referencia para la práctica profesional a la hora de valorar la conveniencia de

los tests existentes en la literatura en relación con los objetivos de la evaluación psicológica.

Muñiz, J. y Fernández-Hermida, J. R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests. *Papeles del Psicólogo*, 31 (1), 108-121. Disponible en línea en http://www.papelesdelpsicologo.es/pdf/1801.pdf.

En este artículo los profesores Muñiz y Fernández-Hermida presentan una parte de los resultados de un estudio dirigido por la European Federation of Psychologists' Associations analizando las opiniones sobre el uso profesional de los tests de psicólogos españoles. Estos resultados, basados en el análisis de las respuestas de 3.126 profesionales de la psicología clínica, educativa y del trabajo, nos dan una interesante fotografía de su actitud general hacia el uso de los tests como instrumentos de evaluación psicológica en España, al tiempo que permite poner de manifiesto algunas limitaciones importantes que deben ser resueltas en el futuro.

Lang, F. (2009). El principio de responsabilidad. *Papeles del Psicólogo*, 30 (3), 220-234. Disponible en http://www.papelesdelpsicologo.es/pdf/1751.pdf.

Este texto forma parte de otro número interesante de la revista Papeles del Psicólogo dedicado a la discusión de las cuestiones éticas y deontológicas vinculadas a la práctica profesional de los psicólogos. En el artículo, el coordinador del Comité de Ética de la European Federation of Psychologists' Associations propone una discusión de la responsabilidad profesional hacia las personas, las comunidades y la sociedad en general de acuerdo con los principios de su Meta-Code of Ethics. Este artículo es una versión ampliada del capítulo con el que Lang contribuyó al libro *Ethics for European psychologists* y da la oportunidad de reflexionar en torno a unos ejemplos concretos que ilustran los diferentes aspectos en los que se concreta el principio general de responsabilidad. Por lo que nos ocupa en esta aproximación a la psicometría, son de especial relevancia los ejemplos 4, 7 y 8, donde plantea algunas consideraciones importantes en relación con las consecuencias derivadas del uso profesional de los tests como instrumentos de evaluación psicológica.

International Test Commission (2000). *International guidelines for test us*. Disponible en línea en http://www.intestcom.org/upload/sitefiles/41.pdf.

En el ámbito de los códigos deontológicos profesionales de los psicólogos, estas directrices suponen una importante contribución desarrollada por la International Test Commission con el objetivo de ofrecer una estructura coherente bajo la cual se puedan entender y aplicar los diferentes códigos y estándares nacionales que desarrollan los aspectos éticos y deontológicos que afectan al uso de tests. Más allá del interés de su articulado para una práctica profesional responsable, estas directrices ponen de manifiesto la importancia del desarrollo y la adquisición de las competencias necesarias para llevar a cabo la administración de tests, la interpretación y comunicación adecuadas de los resultados, y la resolución de las dificultades, malentendidos y conflictos que se puedan derivar. Disponen de una versión traducida al castellano que esta organización proporciona gratuitamente bajo demanda a través de su página web.

Julio Meneses
Vila de Gràcia, junio de 2013

Capítulo I

Aproximación histórica y conceptos básicos de la psicometría

Julio Meneses

Esta introducción a la psicometría se propone dar al lector algunas claves importantes para abordar la complejidad de los conocimientos, procedimientos y valores vinculados con el desarrollo y la administración de tests. Partiendo del encuadre de la psicometría en el contexto general de la psicología, empezaremos desarrollando una aproximación histórica a su nacimiento, desde los antecedentes remotos hasta los primeros desarrollos, describiremos el estatus actual de la psicometría como disciplina científica y, finalmente, construiremos una definición formal teniendo en cuenta el papel que desempeñan los tests como instrumentos de evaluación psicológica. A continuación abordaremos los fundamentos de la psicometría, donde tendremos la oportunidad de ofrecer una definición y clasificación de los tests, abordar los modelos de medida psicométrica, introducir la teoría clásica de los tests y recapitular el proceso de inferencia psicométrica en el que se basan los tests. Finalmente, discutiremos las cuestiones relativas a la construcción y administración de tests, tratando las diferentes fases implicadas en el desarrollo y ofreciendo algunos criterios importantes para la valoración de los tests disponibles en la literatura. Concluiremos esta introducción con una discusión de los aspectos éticos y deontológicos vinculados al uso de tests en la práctica profesional de la psicología.

1. La psicometría en el contexto de la psicología

Entender un campo de estudio determinado implica, en la mayoría de las ocasiones, conocer sus raíces fundacionales. Es decir, conocer los problemas a los que

se enfrentaron los pioneros, la manera como los entendían y las soluciones que les dieron. A pesar de que este no es el lugar para hacer un viaje en profundidad[1], una aproximación histórica a la psicometría debe empezar recogiendo lo que se consideran sus orígenes y antecedentes. A pesar de que el desarrollo y la administración de tests psicológicos es una práctica desarrollada fundamentalmente a partir del siglo XX, es posible encontrar algunos antecedentes remotos en culturas tan antiguas como la china. Partiendo de estos antecedentes, y repasando algunas de las contribuciones más importantes que han contribuido a su desarrollo como disciplina científica, estaremos en disposición de ofrecer una definición formal de la psicometría que nos permita situarla en el contexto general de la psicología.

1.1. Una aproximación histórica a la psicometría

Tal y como se suele señalar, el desarrollo de las primeras dinastías del *antiguo imperio chino* generó los primeros sistemas de evaluación de los individuos en función de su habilidad. A pesar de que algunos cronistas han apuntado a referencias tan antiguas como el año 2000 a. de C. (por ejemplo, podéis ver DuBois, 1970), el estudio de las evidencias arqueológicas ha puesto en entredicho esta antigüedad. En cualquier caso, tal y como ha apuntado Bowman (1989), las pruebas documentales que se conservan permiten situar estos orígenes en un periodo relativamente más reciente, durante la dinastía Tang (618-907). Durante estos años se desarrolló un sistema de evaluación imperial que permitió la selección y promoción de los funcionarios de los diferentes departamentos de la Administración. Este sistema tuvo un importante impulso durante la dinastía Ming (1368-1644), cuando se estableció un examen institucional según el mérito para todos los funcionarios de los diferentes niveles territoriales –desde el nivel municipal al nacional–, poniendo en marcha uno de los primeros sistemas de clasificación oficial mediante la expedición de los primeros títulos formales que acreditaban el acceso a los diferentes niveles de responsabilidad.

1. Entre otros, podéis consultar los trabajos clásicos de Goodenough (1949), DuBois (1970), Hilgard (1987), y más recientemente, Buchanan y Finch (2005), o Jones y Thissen (2007).

Estos sistemas de evaluación son fácilmente homologables a los que la *educación formal europea*, especialmente la universitaria, desarrolló a partir de la introducción de los exámenes orales a sus estudiantes a partir del siglo XIII. Tal y como señala Rogers (1995), la invención y la incorporación del papel a la vida cotidiana a partir del siglo XVI facilitó el tránsito hacia las pruebas escritas que durante el siglo XIX se convirtieron en los primeros mecanismos de selección competitiva de los estudiantes universitarios. Como antecedentes remotos de la psicometría, tanto la evaluación en la educación formal como el establecimiento del sistema de evaluación imperial supusieron un importante cambio en la concepción sobre el juicio de las capacidades de las personas. De este modo, se fue trasladando la confianza en el juicio personal basado en impresiones, hacia la administración de pruebas institucionales basadas en una autoridad imparcial, que objetivaba las destrezas y los conocimientos requeridos en los ámbitos educativo y administrativo.

A pesar de que en algunas ocasiones se ha pasado por alto, otro antecedente remoto se encontraría en los inicios de la *evaluación psiquiátrica* a mediados del siglo XIX. De acuerdo con el trabajo de Bondy (1974), es conveniente tener presente el esfuerzo de los primeros profesionales orientados al estudio de los problemas mentales y las lesiones cerebrales en el establecimiento de lo que se podrían considerar las primeras pruebas de evaluación psicológica. Así, por ejemplo, se desarrollaron los primeros tests para evaluar las consecuencias del daño cerebral, alguno de ellos tan elaborado que exigía una administración durante periodos de 100 horas. A pesar de sus limitaciones, entre otras la ausencia de procedimientos estándares en su uso, estas primeras pruebas incorporaron muchos de los elementos que, con la evolución del estudio de los síntomas psicológicos, todavía son utilizados en las pruebas de diagnóstico actuales.

Desarrolladas las bases para el examen individual, tanto en contextos de actividad cotidiana como en situaciones de trastorno psicológico o accidente, los antecedentes recientes de la psicometría se encontrarían en el desarrollo del *estudio sistemático de las diferencias humanas* durante el siglo XIX. Primero gracias a los trabajos de Friedrich W. Bessel (1784-1846) y Carl F. Gauss (1777-1855) –que fueron pioneros en el estudio de las diferencias individuales en la percepción en el campo de la astronomía– y después a las contribuciones de Gustav T. Fechner (1801-1887) y Hermann von Helmholtz (1821-1894) en el desarrollo de la psicofísica –que supuso el inicio de la psicología como disciplina académica–, ambas aproximaciones constituyeron un avance importante en la sistematización de la medida de las sen-

saciones psicológicas producidas por la estimulación física (Boring, 1978, para una revisión histórica). Así, mediante el desarrollo de las primeras leyes que relacionan estímulos físicos y sensaciones, se fueron asentando las bases para la medida psicológica. De este modo se perfeccionaron los métodos de presentación de los estímulos y el registro de las respuestas, se trabajó en la mejora de la precisión de las medidas y se adoptaron condiciones controladas para su consecución. Todos estos elementos son indispensables para el planteamiento posterior de los problemas más importantes que la psicometría debería afrontar en su desarrollo inicial.

Así, una vez apuntados los antecedentes remotos y más recientes, debemos citar los trabajos de Sir Francis Galton (1822-1911), James McKeen Cattell (1860-1944) y Alfred Binet (1857-1911) como los verdaderos *pioneros de la psicometría moderna*. Como continuadores de las aproximaciones anteriores al estudio sistemático de las diferencias humanas, Galton y Cattell contribuyeron necesariamente al establecimiento de la psicología experimental en Europa y Estados Unidos, respectivamente, mediante la creación de los primeros laboratorios antropométricos para el estudio de las características humanas (Valentine, 1999; Sokal, 1982) y el desarrollo de los primeros tests para la evaluación de las diferencias sensoriales, perceptivas y de comportamiento. Abriendo un nuevo camino para el estudio científico de la psicología, los dos autores se aventuraron a postular relaciones entre sus medidas y el intelecto, llegando a conclusiones a veces controvertidas y otras ampliamente criticadas con posterioridad por su simplismo. Por su parte, Binet adoptó un enfoque innovador y fue responsable de lo que se considera el primer test de aplicación general para la medida de las habilidades cognitivas.

Respondiendo a la necesidad de identificar a los estudiantes con problemas para lograr los objetivos de la instrucción ordinaria en las escuelas de París a principios del siglo XX, Binet y su colega Théodore Simon recogieron los incipientes movimientos en la evaluación de la discapacidad cognitiva mediante baterías de tests (Nicolas y Ferrand, 2002). A partir de ellos, y a petición del Ministerio de Educación francés, desarrollaron en 1905 la primera prueba que permitió clasificar a los niños según su inteligencia (Wolf, 1969). Este trabajo permitió el nacimiento del interés de la psicología hacia la atención a las necesidades educativas especiales, pero no fue hasta 1908 cuando estos autores publicaron una revisión de su escala, que permitió medir lo que denominaron el *nivel mental*. Gracias a la adopción de un grupo de referencia compuesto por niños de entre 3 y 13 años, y una vez ordenados los ítems en función de la edad en la que eran típicamente re-

sueltos, desarrollaron el primer test para cuantificar la inteligencia. Esta cuantificación, tomando como referencia la edad biológica del niño, condujo a la definición de lo que se conoce, hasta la actualidad, como *edad mental*.

Adoptando el mismo enfoque orientado a la práctica, la incipiente psicometría se vio impulsada en los circuitos académicos en torno al debate sobre la *medida de la inteligencia* mediante el desarrollo de tests (Martin, 1997). A partir de los trabajos de Binet en el campo de la educación, el interés por la evaluación de la cognición dio el salto al circuito internacional de la mano de Lewis M. Terman (1877-1956), quien en 1916 publicó la revisión Stanford-Binet. De acuerdo con el relato de Goodenough (1949), esta contribución estableció un importante hito en el desarrollo de los tests tal y como los conocemos en la actualidad. Revisando los ítems originales, incorporando pautas claras de administración y destacando la importancia de la representatividad de las muestras para la correcta interpretación de sus puntuaciones, se convirtió en la escala de referencia para la medida de la inteligencia durante las siguientes décadas. De hecho, fue otro paso importante para el desarrollo de la psicometría, en la medida en que se extendió el uso de esta escala al ejército norteamericano durante la Primera Guerra Mundial. En este contexto, y gracias al apoyo del Gobierno de Estados Unidos, la psicometría creció en el centro de las políticas militares (entre otros, podéis ver Zeidner y Drucker,1988).

Recuperando el espíritu clasificador, los tests deberían servir como instrumentos de evaluación psicológica para el reclutamiento de los nuevos soldados. Así, a partir de la generalización del uso de la revisión Stanford-Binet, se desarrollaron nuevos tests y la administración en grupo para hacer más eficiente el procedimiento de medida. De estos esfuerzos, entre otros, hay que destacar la impronta de Robert Yerkes (1876-1956) en el desarrollo de los tests Army Alpha y Beta (Carson, 1993), que también abordaron una controversia importante sobre los *sesgos culturales de los tests* que se fue construyendo durante estas primeras décadas del siglo XX. Tal y como se ha discutido ampliamente (Jensen, 1980), las nuevas pruebas para medir la inteligencia podían no estar libres de influencias culturales, de manera que infravalorarían a aquellos que no hablaran la lengua inglesa, a los analfabetos y a quienes sufrieran alguna discapacidad visual o auditiva. Recogiendo los incipientes movimientos en torno a las pruebas no verbales, la versión Beta del test de inteligencia del ejército norteamericano supuso el reconocimiento de la importancia de estas diferencias, así como la necesidad de minimizarlas en cualquier contexto en el que se utilicen los tests como instrumentos de evaluación psicológica.

Una cuestión interesante, por las consecuencias que tuvo más adelante en el desarrollo de la psicometría, fue la concepción que se fue construyendo sobre el propio concepto de inteligencia. Arraigada en las nociones biologicistas y hereditarias de los pioneros, y sustentada en los desarrollos estadísticos de Charles Spearman (1863-1945), una importante corriente concibió la inteligencia como un único factor –el factor g, según Spearman–, que explicaría las puntuaciones en los diferentes tests de inteligencia desarrollados hasta el momento. En cambio, desde la posición "alternativa", que encabezó Louis Leon Thurstone (1887-1955), la inteligencia estaría compuesta en realidad por varios factores específicos. Sus avances en el desarrollo estadístico de las *técnicas de análisis factorial* que introdujo Spearman permitieron comprender mejor esta aparente contradicción (por ejemplo, podéis ver Cattell, 1943). De hecho, se podía entender el factor general de inteligencia como un tipo de factor subyacente que recogería la variabilidad de las aptitudes específicas de Thurstone, ya fueran relativas a la comprensión verbal, la competencia numérica o la habilidad espacial, entre otras. Del mismo modo, una vez controlado el efecto del factor general, el planteamiento de unos factores grupales por parte de Spearman permitiría recoger la variabilidad común observada en tests que comparten demandas verbales, numéricas o espaciales similares. Más allá de las consecuencias teóricas sobre la medida de la inteligencia, el perfeccionamiento del análisis factorial permitió pavimentar el camino estadístico para el desarrollo de la psicometría hasta finales del siglo XX.

Además, esta visión de la inteligencia como un conjunto determinado de aptitudes cognitivas específicas contribuyó necesariamente al desarrollo de los *primeros estudios sistemáticos de validez*. Tal y como sucedió durante la Primera Guerra Mundial, el Gobierno norteamericano volvió a incorporar estas prácticas a sus programas de reclutamiento unos años antes de participar en la Segunda Guerra Mundial. A pesar de que los esfuerzos anteriores habían dado sus frutos, el ejército norteamericano encontró en este planteamiento la manera de resolver algunas limitaciones del enfoque unitario de la inteligencia, principalmente por su escasa capacidad para seleccionar candidatos con perfiles muy especializados. Fue durante este periodo cuando, bajo la dirección de John C. Flanagan (1906-1996), las fuerzas de aviación administraron un conjunto de baterías de tests para seleccionar y clasificar a los pilotos, ingenieros de vuelo y otros técnicos encargados de los instrumentos de navegación. El estudio de la relación entre las puntuaciones en las diferentes aptitudes y el éxito en la formación posterior de los

reclutas fue clave en este proceso (Goslin, 1963), permitiendo la mejora de los procedimientos de selección y clasificación, y conformando uno de los primeros estudios empíricos de la validez de las medidas psicológicas.

Sin embargo, no debemos olvidar el trabajo que paralelamente se llevó a cabo en relación con la *evaluación de la personalidad*. Todavía en el contexto de las políticas militares de la Primera Guerra Mundial, el Gobierno de Estados Unidos se enfrentó a otro tipo de problema práctico. Más allá de la selección y clasificación de los reclutas, un importante esfuerzo se dirigió a la identificación de candidatos susceptibles de sufrir trastornos psicológicos. Para minimizar su presencia en sus filas, Robert S. Woodworth (1869-1962) fue el encargado de desarrollar un nuevo tipo de prueba que permitiera evaluar la estabilidad emocional de los soldados. Este test, su *Personal Data Sheet*, introdujo un conjunto de preguntas con respuesta positiva o negativa que, a diferencia de la evaluación de la inteligencia, no contenía respuestas necesariamente correctas o incorrectas. De este modo, Woodworth desarrolló una prueba que permitía detectar los casos problemáticos no ya mediante la comparación de la respuesta individual con una muestra representativa de la población, sino a partir de las respuestas dadas por individuos con trastornos psicológicos ya diagnosticados. Este trabajo asentó las bases para el desarrollo de los tests de evaluación de la personalidad posteriores (Gibby y Zickar, 2008), así como empezó a definir las bases para tratar adecuadamente la posibilidad de fraude en la respuesta, tanto para ocultar como para simular que se sufre un trastorno.

El prolífico Thurstone, que desarrolló las primeras pruebas de análisis de consistencia interna de los tests en el ámbito de la evaluación de la personalidad, hizo otra contribución importante al desarrollo de la psicometría. Partiendo de los experimentos desarrollados por los pioneros (Gulliksen, 1968), especialmente en la rama de la psicofísica, propuso en 1927 la ley del juicio comparativo como el método para la *medida de las actitudes, las preferencias y los valores*. Esta innovación consistió en trasladar el juicio perceptivo sobre parejas de estímulos físicos –por ejemplo, la elección entre dos sonidos según su intensidad– a la valoración de características psicológicas no estrictamente relacionadas con las propiedades físicas –por ejemplo, la elección entre dos comportamientos según su aceptabilidad. Después de proponer una segunda ley, la ley del juicio categórico, la aproximación de Thurstone a la medida de las actitudes estaba preparada para ser aplicada al juicio general, es decir, sin requerir la comparación entre parejas de fenómenos. De este modo, tal y como pretendían Thurstone y sus continuadores (Bock y Jones, 1968), la psicome-

tría dispondría de un método de escalamiento para medir y tratar numéricamente valoraciones individuales subjetivas en la búsqueda de una representación objetiva de los fenómenos psicológicos. Este método dio un soporte metodológico importante en las décadas siguientes de investigación en psicología social.

1.2. La psicometría hoy

Pese a la dificultad de llevar a cabo una aproximación histórica a partir de la selección de algunas de las aportaciones más importantes, este ejercicio nos permite ilustrar el nacimiento de una disciplina a partir de los problemas y las diferentes aproximaciones de los pioneros. La psicometría, un nuevo espacio de trabajo metodológico en torno al desarrollo y a la administración de tests, se fue conformando con un componente aplicado, orientada a las demandas en diferentes contextos, y formó parte de algunos de los debates teóricos más importantes de la propia psicología. Sin embargo, hasta la década de los años treinta del siglo XX no podemos situar los *inicios de su constitución como disciplina científica*[2] tal y como la conocemos en la actualidad. Es importante empezar señalando de nuevo el papel decisivo de Thurstone, quien, con el objetivo de establecer y promover la psicología como una ciencia cuantitativa (Samejima, 2000), en 1935 fundó y fue el primer presidente de la Psychometric Society. Además, fue el impulsor de la primera revista especializada que todavía es de referencia obligada, la revista *Psychometrika*, que publicó algunos de los trabajos más importantes sobre los que se formalizó la psicometría. Pocos años después, en 1946, fue también el primer presidente de la División de Evaluación y Medida de la American Psychological Association.

En este sentido, podríamos definir el periodo entre los años treinta y sesenta como la *época dorada de la psicometría*. En este periodo es cuando se publican, además, los libros y manuales más importantes que la vertebrarían. Entre los manuales hay que destacar *The reliability and validity of tests*, de Thurstone (1931), que sistematizaba lo que se había desarrollado hasta el momento en relación con la teoría de los tests y sugería el papel central de la fiabilidad como

2. El lector interesado en profundizar sobre la constitución de la psicometría como disciplina científica desde una perspectiva histórica puede consultar las revisiones de Hambleton (1994), Brennan (1997), Traub (1997) o Bock (1997).

requisito para la validez de las medidas en la psicometría. Volveremos más adelante a estas cuestiones, elementos básicos de la medida indirecta de los fenómenos psicológicos mediante tests. De manera contemporánea, fue también importante la primera edición del manual *Psychometric methods* de Guilford (1936), un intento de organizar el campo propio de la psicometría en torno a la teoría de los tests, el escalamiento psicológico y el psicofísico. La teoría clásica de los tests, como fue ampliamente conocida a partir de los trabajos de Spearman sobre la estimación de los errores de medida, empezaba su camino en los circuitos docentes universitarios encargados de formar a los futuros psicólogos.

Por otro lado, durante los años treinta, y especialmente en los cuarenta, uno de los debates más apasionantes abiertos por la psicometría tuvo lugar sobre el propio concepto de *medida*. Inspirados por la British Association for the Advancement of Science, psicólogos y físicos se propusieron el reto de decidir si, con los avances en el desarrollo de los tests, la medida psicológica podía ser encuadrada dentro del modelo general de medida de los atributos físicos. En síntesis, medir consistía en cuantificar, es decir, determinar la magnitud con la que un atributo está presente en un objeto. Y para hacerlo, la medida de los atributos físicos dependía de la capacidad de observación de relaciones entre objetos como consecuencia de una operación o manipulación empírica. Por cuanto que esta posibilidad solo podía ser satisfecha con los objetos físicos, la controversia tuvo una respuesta inicial contundente. Tal y como concluyeron Ferguson y sus colaboradores (1940), la medida psicológica que se proponía la psicometría simplemente no era posible, puesto que este modelo no se podía extender para incluir los atributos psicológicos que, por definición, no eran observables ni manipulables empíricamente. El debate, sin embargo, no quedó cerrado y fue Stevens (1946), con su trabajo *On the theory of scales of measurement*, quien dio un paso fundamental para el desarrollo de lo que posteriormente fue denominado el modelo representacional de medida psicométrica.

Stevens definió la medida como el proceso de asignación de números a objetos o acontecimientos de acuerdo con unas reglas, producto de las cuales se obtendrían los diferentes tipos de escalas propuestas: nominal, ordinal, de intervalo y de razón. Medir, en este sentido, no consistiría únicamente en cuantificar, sino que sería el producto de la utilización de diferentes reglas que, en último término, determinarían el tipo de operaciones –o técnicas– estadísticas permitidas en cada escala. No exenta de críticas, es importante resaltar la importancia de su contri-

bución como un primer intento para superar las restricciones impuestas por la cuantificación de los atributos físicos con el objetivo de resolver la controversia en torno a la medida de los fenómenos psicológicos. Así, el debate sobre los modelos o paradigmas de medida formó parte importante de la agenda psicométrica y se pusieron las bases para el desarrollo posterior de otros modelos, como el operacional y el clásico, sobre los que volveremos más adelante. Por otro lado, cerrando este breve repaso a las contribuciones importantes durante los años cuarenta, no podemos dejar de hacer mención a la publicación por parte de Thurstone (1947) del influyente *Multiple factor analysis*, que a partir de los trabajos de Spearman, Kelley y Burt proporciona el soporte estadístico necesario para la construcción y validación de los tests durante las siguientes décadas.

En la década de los cincuenta asistimos a la publicación de las otras dos obras de referencia para la psicometría moderna. Instalada en el circuito académico, el trabajo llevado a cabo por varios autores converge en lo que se ha denominado teoría clásica de los tests (TCT). Y lo hace a manos de Gulliksen (1950) en su *Theory of mental tests*, donde formaliza el modelo lineal clásico por primera vez y define sus asunciones principales. Como teoría de los tests, la TCT propone un nuevo enfoque basado en el concepto de puntuación verdadera. Partiendo de la puntuación empírica obtenida mediante los tests, y siguiendo un conjunto determinado de supuestos, el objetivo es descomponerla en dos partes fundamentales para valorar el error asociado al proceso de medida y, así, inferir el valor real que se pretende medir. No sin dificultades se convirtió en poco tiempo en la teoría de los tests de referencia y estimuló tanto el debate sobre la propia medida psicológica como el proceso de desarrollo de los tests durante las siguientes décadas. De manera análoga, Torgerson (1958) publica *Theory and methods of scaling*, y establece el canon para el escalamiento psicofísico y psicológico, es decir, la ordenación de los estímulos de manera paralela al de las personas.

Finalmente, los años sesenta son una década de *revisión crítica y apertura de nuevos caminos* para la psicometría. A partir de la TCT, el debate sobre la estimación de los errores cristalizó en dos nuevas corrientes en el desarrollo de tests: la teoría de la generalizabilidad y la teoría de respuesta al ítem. Para incrementar la precisión de las medidas, Cronbach y sus colaboradores (Cronbach, Rajaratnam y Gleser, 1963; Gleser, Cronbach y Rajaratnam, 1965). propusieron la teoría de la generalizabilidad, que, mediante la aplicación del análisis de varianza, permitiría descomponer el error genérico propuesto por la TCT en la búsqueda

de sus diferentes elementos. Así, de acuerdo con esta teoría, el análisis de la fiabilidad se basa en el diseño de investigaciones que permiten analizar las diferentes fuentes de error –facetas, según sus términos– que afectarían al proceso de medida. Entre estas, por ejemplo, la forma del test, los ítems que lo componen, las ocasiones en las que se administra o los participantes. A pesar de que fue una importante innovación, la complejidad de esta teoría limitó su difusión a la práctica y en poco tiempo sus avances fueron básicamente reformulados por un segundo enfoque, la teoría de respuesta al ítem (TRI).

Así, la TRI se presenta como la respuesta a las críticas principales que había recibido la TCT. En este sentido, es conveniente hacer mención a la publicación por parte de Lord y Novick (1968) de su *Statistical theories of mental test scores*. En este trabajo, también con espíritu de síntesis y análisis crítico del trabajo desarrollado en el contexto de la TCT, se plantean las dificultades que este modelo no estaba siendo capaz de resolver. Básicamente, los problemas principales podrían ser organizados en torno a la dependencia de los instrumentos y las muestras utilizadas en el proceso de construcción y administración de los tests. Durante las décadas posteriores la TRI ocupará, junto a la TCT, un lugar privilegiado entre las teorías de los tests desarrolladas por la psicometría. Volveremos más adelante sobre estas cuestiones, pero vale la pena cerrar esta etapa dorada de la psicometría moderna señalando un último desarrollo que también tuvo sus inicios en la década de los sesenta. Se trata de los tests referidos a criterio, que, en el contexto de la educación, tienen como objetivo evaluar la destreza de las personas en un campo de conocimiento muy bien delimitado.

Evaluar en función de criterios y no de la norma –es decir, escalar u ordenar a los individuos comparando sus puntuaciones– no es un enfoque nuevo. Al contrario, como hemos visto en el recorrido histórico del nacimiento y establecimiento de la psicometría como disciplina científica, este modo de evaluación ha sido objeto de interés con anterioridad, pero gracias a las contribuciones de Glaser (1963) y Popham y Husek (1969) el trabajo en este contexto queda formalizado. Lo que es innovador, desde el punto de vista del desarrollo de los tests, es el enfoque en el procedimiento para establecer los estándares de evaluación, así como en la consistencia y precisión con la que se clasifica a los individuos de acuerdo con estos estándares. A pesar de que nació en el marco de la TCT, aplicaciones posteriores de la TRI han permitido aprovechar los avances en el desarrollo de los tests referidos a criterio como instrumentos de medida de los fenómenos psicológicos.

1.3. La psicometría en el contexto de la evaluación psicológica

Una vez recorrido el camino de los antecedentes y los orígenes de la psicometría moderna, estamos en disposición de abordar formalmente su definición. Y podemos hacerlo a partir de las definiciones que se han ido proponiendo en los textos de referencia. Sin embargo, esta tarea no está exenta de riesgo. Los autores elaboran definiciones concretas que, probablemente, no reflejan todos los matices con los que después las contextualizan en sus textos. A pesar de esto, podemos aprovechar estas propuestas para valorar los diferentes aspectos en los que se pone un énfasis especial. Al fin y al cabo, no cabe todo en una definición.

Más allá del tópico que dice que hay tantas definiciones como autores, podemos distinguir tres grandes aproximaciones. Por un lado, un primer grupo de definiciones aborda la psicometría a partir de los *instrumentos que utiliza*. Desde esta corriente se circunscribe la psicometría como la disciplina encargada de desarrollar los fundamentos para la construcción y administración de tests (por ejemplo, ver Martínez Arias, Hernández Lloreda y Hernández Lloreda, 2006). Esta definición, a pesar de que corta y concisa, presenta algunas dificultades. En primer lugar, la palabra *test* es polisémica, por lo que según el contexto en el que sea empleada puede ser sinónimo de otros términos con significados diferentes, como *prueba*, *examen* o incluso *ensayo*. En segundo lugar, más allá de la precisión semántica que se puede sobreentender en el contexto de la psicología, no deja de ser en cierto modo restrictiva. Si bien es cierto que los tests psicológicos son los instrumentos específicos que desarrollan y administran los psicómetras, también lo es que son el resultado aplicado de un proceso más amplio caracterizado por el desarrollo de teorías y métodos orientados a la medida indirecta de los fenómenos psicológicos.

Una segunda corriente define la psicometría a partir del *objeto al que somete o aplica su interés*. Así, la psicometría se define de manera más o menos genérica como la disciplina científica orientada a la evaluación o medida de los fenómenos psicológicos (por ejemplo, ver Rust y Golombok, 2009). Esta definición se ajusta de hecho al significado etimológico de la palabra *psicometría*, que en sus orígenes griegos podemos encontrar en la yuxtaposición de las palabras *psique* –que significa 'alma', 'aliento' o 'intelecto'– y *metron* –en referencia al proceso de medida. En esta línea se puede situar la definición de Kline (1998), que aborda específicamente la psicometría como la tarea de desarrollar medidas científi-

cas fundamentales en las áreas de la personalidad y las capacidades. En este contexto, la cientificidad se convierte en sinónimo de *estandarización* y, a pesar de no hacer referencia explícita, alude indirectamente al uso de tests en el contexto del método científico. Otro implícito importante es, con relación al estatus métrico de las puntuaciones obtenidas como resultado del proceso de medida psicológica, el carácter cuantitativo que debería tener la psicometría como disciplina científica. Como ya hemos avanzado en la aproximación histórica, este es un tema controvertido al que deberemos volver más adelante.

En una posición intermedia, un tercer grupo de autores sitúa la psicometría en la *intersección de las dos corrientes anteriores*. En esta línea podemos señalar por ejemplo la definición de Buchanan y Finch (2005), que especifica los dominios de medida estandarizada mediante tests en relación con las habilidades, los atributos y las características psicológicas. Es importante señalar que, a pesar de que en algunos casos se hace referencia explícita al uso de tests, entre estas definiciones podemos encontrar también un reconocimiento explícito del sentido más amplio del proceso de medida. Así, Holden (2000) define la psicometría como la teoría y la técnica de medida que, en el contexto de la psicología, se encarga de los factores que son medibles. En este sentido, Muñiz (2003) lo hace definiendo la psicometría como el conjunto de métodos, técnicas y teorías implicadas en la medida de las variables psicológicas, teniendo en cuenta su especialización en las propiedades métricas exigibles a este tipo de medida. Finalmente, Jones y Thissen (2007) describen la psicometría como la disciplina –cuantitativa, especifican en este caso– encargada de desarrollar modelos y métodos orientados principalmente al resumen, la descripción y el establecimiento de inferencias a partir de los datos recogidos en la investigación psicológica.

Como hemos podido observar, la definición de la psicometría se ha desarrollado fundamentalmente en torno a los medios o instrumentos que utiliza –los tests– y el objeto que persigue con su uso –la medida de los fenómenos psicológicos. En cambio, una aproximación comprensiva a esta definición debería considerar un tercer elemento importante vinculado a la *finalidad a la que sirve*. Tal y como hemos denominado este apartado, la psicometría adquiere todo su significado en relación con el área de la psicología en la que desempeña su papel fundamental: la evaluación psicológica. De este modo, hemos de tener presente que la medida de los fenómenos psicológicos mediante el desarrollo y la admi-

nistración de tests no es más que una parte del proceso general de evaluación, y no siempre es la más importante (Fernández-Ballesteros, 1997).

La psicometría, en este sentido, contribuye al desarrollo de la evaluación psicológica proporcionando teorías, métodos y técnicas que, en última instancia, permiten describir, clasificar, diagnosticar, explicar o predecir los fenómenos psicológicos objeto de medida. De hecho, contextualizada así la psicometría podemos ir un paso más allá y señalar que, a su vez, la evaluación psicológica no se encuentra en ningún otro lugar que al servicio de la intervención psicológica. Es decir, la finalidad última que guía el desarrollo y la administración de tests es la de contribuir a la recogida de las evidencias necesarias que permitan a los psicólogos tomar una decisión u orientar alguna acción. Evidentemente, esta intervención dependerá del contexto en el que la evaluación haya sido desarrollada, pero es importante tenerlo presente para entender no solo qué es la psicometría, sino también su importancia para el ejercicio profesional de los psicólogos en los diferentes contextos en los que intervienen.

Partiendo de este enfoque comprensivo a la psicometría, podemos sintetizar las diferentes aproximaciones y formular finalmente una definición que incorpore su papel en el marco de la evaluación y la intervención psicológicas.

Así, la psicometría es una rama de la psicología que, mediante teorías, métodos y técnicas vinculados al desarrollo y la administración de tests, se ocupa de la medida indirecta de los fenómenos psicológicos con el objetivo de hacer descripciones, clasificaciones, diagnósticos, explicaciones o predicciones que permitan orientar una acción o tomar decisiones sobre el comportamiento de las personas en el ejercicio profesional de la psicología.

Para concluir este ejercicio de definición formal, es importante discutir brevemente las diferencias que existen entre la psicometría y otra disciplina también interesada por la medida psicológica: la *psicología matemática*[3]. En este sentido, podemos reconocer la psicología matemática como la rama de la psico-

3. Podéis ver, entre otros, Jáñez (1989) y Padilla, Merino, Rodríguez-Miñón y Moreno (1996).

logía interesada en el desarrollo de modelos de los procesos perceptivos, cognitivos y motores con el objetivo de establecer leyes que relacionen estímulos físicos con los comportamientos a partir de estudios casi exclusivamente experimentales. Son muchos los autores comunes a las dos orientaciones desde los orígenes del estudio sistemático de las diferencias humanas en el siglo XIX, pero, a diferencia de la psicometría, la psicología matemática no está tan interesada por las diferencias individuales como en la definición de leyes generales que modelen el comportamiento medio de las personas.

2. Fundamentos de la psicometría

Una vez abordada la psicometría desde una perspectiva histórica, tratando sus raíces y su construcción como disciplina científica en el contexto de la psicología, en este segundo apartado revisaremos sus fundamentos. Empezaremos discutiendo una definición y clasificación de los tests, progresaremos por los diferentes modelos de medida psicométrica desarrollados en las diferentes teorías de los tests y realizaremos una breve introducción a la teoría más extendida en la práctica actual: la teoría clásica de los tests. Por último, recapitularemos estos fundamentos atendiendo al proceso que la psicometría sigue para establecer sus inferencias sobre los fenómenos psicológicos no observables a partir de las puntuaciones obtenidas mediante tests. Empecemos, pues, con una definición y clasificación de los tests.

2.1. Definición y clasificación de los tests

De acuerdo con la definición del manual clásico de Anastasi (1988), un test psicológico (en adelante, diremos un test) es un procedimiento de medida objetiva y estandarizada de una muestra de comportamientos. Otras definiciones son posibles[4] pero esta contiene tres elementos fundamentales que nos permi-

4. Podéis ver, entre otros, Kaplan y Saccuzzo (2001), Murphy y Davidshofer (2005) o Urbina (2004).

ten abordar sistemáticamente las características más importantes que cumplen de manera genérica los tests. De un modo u otro, ya las hemos ido introduciendo en este texto.

En primer lugar, la medida mediante el desarrollo y la administración de tests es, o al menos pretende ser, *objetiva*. En este sentido, la objetividad hace referencia a la sustitución del juicio personal basado en criterios subjetivos por un conjunto de normas determinadas y conocidas que permiten obtener e interpretar las puntuaciones de los individuos en igualdad de condiciones. Asimismo, la medida que pretenden los tests es *estandarizada*, en cuanto que las puntuaciones obtenidas dependen de un procedimiento establecido de administración, corrección e interpretación que las hacen, o las deberían hacer, invariantes del profesional que administra los tests, las condiciones específicas en las que lo hace y el modo como obtiene e interpreta las puntuaciones resultantes de la media. Finalmente, los tests se enfrentan a los fenómenos psicológicos no observables mediante una *muestra de comportamientos*. En la medida en que esta muestra sea representativa del conjunto, las puntuaciones obtenidas permitirán a los profesionales establecer adecuadamente sus inferencias sobre el comportamiento general de las personas más allá de los elementos específicos evaluados mediante tests.

Ya hemos discutido la naturaleza no observable de los fenómenos psicológicos y la dificultad intrínseca que representa para la *medida indirecta mediante tests* que se propone la psicometría. Pero es importante resaltar esta cuestión cuando la comparamos con otros tipos de medida científica. En este sentido, la medida de los fenómenos psicológicos no resulta tan sencilla como la de las magnitudes físicas –por ejemplo, el peso o la longitud– de los objetos que las ciencias naturales pueden observar y manipular directamente. Además, tal y como señala García Cueto (1993), existen otras *diferencias importantes* entre este tipo de medida y la medida indirecta de los fenómenos psicológicos que deben ser tenidas en cuenta.

Por un lado, el objetivo habitual de la medida de las magnitudes físicas es obtener información sobre un único objeto. La psicometría, en cambio, se propone desarrollar instrumentos que permitan obtenerla sobre un grupo de individuos con el objetivo de extraer conclusiones sobre cada uno de ellos, sobre el grupo entero, e incluso extrapolar sus resultados a las poblaciones de referencia de donde provienen estos individuos. Por otro lado, la medida de magnitudes físicas parte

de la posibilidad de repetir su procedimiento tantas veces como sea deseado sin variar las condiciones en las que se lleva a cabo la medida. Este procedimiento es incompatible con la medida mediante tests, en cuanto que la repetición en la aplicación de una misma prueba sobre los mismos individuos produce variaciones en las puntuaciones, que pueden ser explicadas por el cansancio o por el efecto de la práctica en el aprendizaje y no por variaciones sustanciales en los fenómenos de interés. A pesar de esto, podemos encontrar algunas prácticas desarrolladas en los contextos aplicados de las ciencias naturales que tienen mucho en común con la administración de tests.

Así, cuando se pretende analizar la contaminación atmosférica de una ciudad o el nivel de alcohol en sangre de una persona, los científicos desarrollan un conjunto de pruebas objetivas –llamados *reactivos*– para aplicarlas de manera estandarizada sobre una muestra representativa del objeto que pretenden medir –ya sea el aire de la ciudad o la sangre del individuo, siguiendo los mismos ejemplos. En este sentido, el proceso de desarrollo y administración de tests es comparable con estas prácticas, y pone de relieve el sentido que la psicometría atribuye a los tests como sus instrumentos de medida. Un test, lejos del dominio de la psicología, no es más que una prueba, un examen o un ensayo. En definitiva, es un *reactivo* que, en el caso de la medida indirecta de los fenómenos psicológicos mediante tests, no tiene otra finalidad que la de producir una reacción en el comportamiento de los individuos para registrarlo y obtener una puntuación como resultado del proceso de medida. Este sentido etimológico de la palabra *test* lo podemos encontrar en la raíz de su definición en el contexto de la psicometría (por ejemplo, Yela, 1984), donde se asimila directamente con el reactivo de las ciencias naturales y se espera de él que, aplicado a un individuo, revele y dé testimonio fiel de los fenómenos psicológicos no observables que son objeto de medida.

Desde este punto de vista, el valor de un test se encuentra en su capacidad para *suscitar y medir comportamientos* que resulten un buen indicador –es decir, una buena representación– del conjunto global de comportamientos implicados en los fenómenos de interés. En este sentido, debemos hacer una precisión importante con relación a los fenómenos que son objeto de medida mediante tests. En este texto hablamos de los fenómenos psicológicos de una manera general y lo hacemos en cuanto que el objetivo de un test es la medida objetiva y estandarizada de una muestra de comportamientos. Esto no significa, sin em-

bargo, que los tests deban dirigirse exclusivamente a elementos con una larga tradición en la teoría psicológica, como son, entre otros, la inteligencia o la personalidad. Por ejemplo, algunos tests se dirigen hacia la medida de las respuestas fisiológicas, como pueden ser la conductividad de la piel o la frecuencia cardíaca. Otros se fijan en cuestiones como las actitudes racistas, el consumo de sustancias o la sociabilidad de las personas. Finalmente, en el campo de la educación, no son pocos los tests cuyo objetivo es evaluar el dominio de los estudiantes en diferentes áreas de conocimiento. En todos los casos hablamos de fenómenos psicológicos en cuanto que responden a objetos de interés para las ciencias sociales y del comportamiento y, además, recaen en los dominios de la medida indirecta mediante tests siempre que se reúnan las condiciones que establece la psicometría.

Así, la medida indirecta que se proponen los tests empieza con una definición precisa de los fenómenos objeto de interés y se sustenta fundamentalmente en dos tipos de teorías. Por un lado, en una *teoría sustantiva* sólida y bien establecida sobre el comportamiento de las personas –por ejemplo, una teoría de la inteligencia–, que dará el soporte teórico necesario para definir los elementos críticos que conforman estos fenómenos y determinará los comportamientos implicados que serán empleados como evidencias observables en el proceso de medida. Por otro lado, en una *teoría de los tests*, que permite establecer las inferencias sobre los fenómenos psicológicos no observables a partir de las puntuaciones obtenidas en el proceso de medida. En su centro, un *modelo de medida* determinado que sirve al propósito de relacionar las puntuaciones obtenidas y los fenómenos objeto de medida, a partir del cual se articula el *proceso de inferencia psicométrica*. Tanto los modelos de medida como el propio proceso de inferencia psicométrica serán tratados en detalle más adelante. Pero es importante hacer mención a ello en este punto porque resultan clave para entender la medida indirecta que se proponen los tests y que los define en relación con otros tipos de medida científica.

Finalmente, una última consideración sobre los tests y los diferentes términos con los que se hace referencia a ellos nos lleva a fijarnos en dos grandes maneras de tratar las respuestas obtenidas para medir los fenómenos psicológicos de interés. Por un lado, los denominados *tests de habilidad o de potencia* tienen como objetivo evaluar la competencia, la aptitud o los conocimientos de los individuos a partir del acierto o la calidad de sus respuestas. Son pruebas que dis-

criminan respuestas correctas e incorrectas y esta es la base para puntuar las ejecuciones individuales. Por otro lado, los denominados *tests de personalidad*[5] tienen un objetivo diferente y pretenden conocer de manera general las motivaciones, preferencias, opiniones o actitudes de los individuos frente a un determinado estímulo. Este segundo tipo de tests no tienen respuestas correctas y por lo tanto no sirven para evaluar el acierto o el error de los individuos. Esta distinción es relevante desde un punto de vista terminológico, dado que los tests de personalidad son muchas veces llamados *cuestionarios*, *inventarios* o *escalas*, a pesar de que este uso no es siempre consistente. De hecho, el término *escala* puede resultar sinónimo de un test cuando está conformado por diferentes partes, o incluso de cada una de estas partes para reflejar las dimensiones o características específicas que miden en el contexto del test global. Por otro lado, otro término también extendido es el de *batería*, que tampoco tiene un significado unívoco y puede hacer referencia tanto a un test compuesto por varias partes, como a una selección de diferentes tests administrados conjuntamente por un profesional en una evaluación psicológica determinada.

Después de esta discusión terminológica, y a pesar de la enorme variedad de tests existente en la actualidad, estamos en disposición de clasificarlos en función de algunas de sus características más importantes. Así:

– **Según el propósito**. Los tests pueden tener diferentes finalidades, y entre ellas podemos destacar dos fundamentales: la *diagnosis*, orientada a la evaluación de las condiciones actuales de los individuos, y la *predicción*, que se propone relacionar la medida actual con el comportamiento de las personas en situaciones futuras. En este sentido, por ejemplo, una cosa es obtener información sobre la destreza de una persona en la resolución de problemas y otra es utilizar esta información para predecir su desempeño en un trabajo determinado. Otros propósitos más específicos son también posibles, como ya hemos comentado en la discusión de la definición formal de la psicometría.

5. A pesar de que el término *personalidad* puede hacer referencia a la rama de la psicología interesada por las características personales que influyen en sus cogniciones, emociones, motivaciones y comportamientos, en este caso se utiliza en un sentido más amplio para hacer referencia a los tests que, en oposición a los de habilidad o de potencia, no pretenden medir las capacidades de las personas.

- **Según el contenido**. De manera general, podemos clasificar los tests según el área de la psicología a la que pertenecen los fenómenos psicológicos que pretenden medir. Esta clasificación, sin embargo, varía según los autores que las realizan, en cuanto que pueden fijarse en diferentes niveles de complejidad al jerarquizar estas áreas. En el nivel más general podemos distinguir básicamente tres grandes grupos, que incluirían los tests orientados a la evaluación de las *habilidades cognitivas, la personalidad y las actitudes*.
- **Según el formato**. Teniendo en cuenta los materiales utilizados es frecuente distinguir los tests de lápiz y papel, de manipulación y de medidas fisiológicas. Nos referimos a los *tests de lápiz y papel* cuando presentan los ítems o preguntas en papel y requieren que el individuo dé algún tipo de respuesta escrita. Los *tests computerizados*, a pesar de que no usan el papel, podrían ser incluidos en esta categoría en cuanto que exigen algún tipo de respuesta escrita mediante dispositivos electrónicos. Por otro lado, son *tests de manipulación* aquellos que presentan una serie de objetos, imágenes o rompecabezas que los individuos deben resolver para demostrar su habilidad. Finalmente, los *tests de medidas fisiológicas* utilizan sensores de distinto tipo para registrar las reacciones de los individuos ante los estímulos físicos presentados.
- **Según el tipo de administración**. Otra manera de clasificar los tests puede tener en cuenta el modo como son administrados y nos permite distinguir los *tests individuales y grupales*. Esta clasificación se solapa con la que hemos hecho según el formato, pero nos permite distinguir tests que requieren una administración a un único individuo o permiten hacerlo con un grupo de individuos al mismo tiempo. También podemos distinguir los *tests verbales y no verbales*, en función de si se hace o no una presentación oral o escrita del test y sus instrucciones.
- **Según el tratamiento de las respuestas**. Siguiendo la diferencia que establecíamos entre los *tests de habilidad o de potencia* y los *tests de personalidad*, podemos clasificar también los tests en dos grandes grupos, en función de si evalúan el acierto en las respuestas para determinar la competencia, la aptitud o los conocimientos de los individuos, o si tratan de evaluar motivaciones, preferencias, opiniones o actitudes. Estos dos tipos de pruebas son también conocidas como *pruebas de ejecución máxima* y *pruebas de ejecución típica*, respectivamente.

– **Según la interpretación de las puntuaciones**. Tal y como ya hemos comentado, las puntuaciones obtenidas mediante tests pueden ser interpretadas según la norma o en referencia a un criterio. Así, los *tests normativos* permiten comparar la puntuación del individuo con la ejecución observada en un grupo de referencia que previamente ha respondido al mismo test. Este grupo puede estar compuesto por muestreo probabilístico, cuando representa a la población de referencia, o no probabilístico, cuando estamos interesados en comparar las puntuaciones con un grupo de personas que cumple unas características determinadas. Por otro lado, los *tests referidos a criterio* toman como referencia la definición de un dominio de conocimientos o habilidades específicos y permiten medir la ejecución del individuo no ya en comparación con un grupo de referencia, sino en función de su grado de adecuación o consecución de este criterio.
– **Según el estatus comercial**. Finalmente, los tests se pueden clasificar en función de si son propietarios o no. Los *tests propietarios o comerciales* son tests que requieren el pago para su uso e incluyen los ejemplares del propio test y un manual que contiene información sobre su desarrollo y sus propiedades psicométricas, las hojas de corrección de las respuestas y las tablas para interpretar las puntuaciones en relación con los grupos de referencia. En otros casos hablamos de *tests abiertos o no comerciales*, y generalmente se pueden obtener contactando con el autor, que normalmente también ha desarrollado algún tipo de manual que contextualiza el test. En algunas ocasiones esta información queda reducida a la publicación de un artículo científico en el que el autor lo introduce, realiza una primera administración y presenta sus propiedades psicométricas.

2.2. Modelos de medida psicométrica

Todo tipo de medida científica, y la medida indirecta de los fenómenos psicológicos mediante tests no es una excepción, se fundamenta en una *definición del propio concepto de medida*. En este sentido, tal y como ya hemos introducido, en el núcleo de las teorías de los tests podemos encontrar un conjunto de modelos que establecen qué es exactamente medir y, como consecuencia, especifi-

can la relación existente entre las puntuaciones empíricas obtenidas y los fenómenos psicológicos no observables objeto de interés. Son los *modelos o paradigmas de medida* que la psicometría ha desarrollado de manera formal a partir del debate iniciado en los años cuarenta sobre la viabilidad de la medida de los fenómenos psicológicos. No obstante, antes de caracterizar estos modelos retomaremos la controversia a partir de las conclusiones de Ferguson y sus colaboradores (1940) para la British Association for the Advancement of Science. Como veremos a continuación, suponen un importante punto de partida a partir del cual la psicometría ha desarrollado sus propios modelos de medida para abordar específicamente los fenómenos psicológicos no observables.

De acuerdo con el trabajo de Campbell (1920), físico y miembro de la comisión que elaboró este informe, el *modelo general de medida desarrollado en la física* se basaría en la existencia de una equivalencia –un isomorfismo, en sus palabras– entre algunos atributos físicos y las propiedades aditivas de los números. Medir consistiría en asignar números a estos atributos para representar la aditividad de los objetos físicos y esta representación solo sería posible en cuanto que se satisficieran dos condiciones: que el atributo se pudiera ordenar, es decir, que fuera posible determinar que el atributo presente en un objeto es menor, igual o mayor que otro; y que se pudiera demostrar empíricamente la existencia de la aditividad mediante la manipulación empírica. Es lo que denominó *medida fundamental*, definiendo la medida de los atributos físicos por analogía con los números a partir de las operaciones llevadas a cabo sobre los objetos para cuantificar sus atributos.

> En un ejemplo sencillo, el peso de dos o más objetos puede ser observado empíricamente mediante una balanza para ordenarlos de menor a mayor, así como se puede demostrar empíricamente que el peso obtenido después de poner estos objetos juntos en la balanza da como resultado la suma de los pesos parciales.

Campbell (1928) extendió este modelo general a lo que denominó como *medidas derivadas*, donde la cuantificación de ciertos atributos físicos no dependería de la manipulación empírica sino del descubrimiento de relaciones matemáticas entre dos medidas fundamentales. Por ejemplo, sería el caso de la densidad de los objetos, medida derivada que respondería a una razón entre su masa y su volumen. En cualquier caso, la clave de la medida para el modelo de Campbell se encontraría en la necesidad de observar y manipular empíricamen-

te los objetos, bien para cuantificar un atributo físico mediante su observación directa, bien para establecer funciones matemáticas que relacionen otros atributos observables. En la medida en que los fenómenos psicológicos no pueden ser observados ni manipulados empíricamente, ni son producto de una relación entre otros atributos observables, la medida indirecta que persigue la psicometría no sería equiparable a la medida fundamental o derivada de la física y, por lo tanto, resultaría inviable. Esta fue la conclusión de la comisión organizada por la British Association for the Advancement of Science, que, como ya hemos comentado, lejos de cerrar la controversia se convirtió en uno de los debates más importantes para la medida indirecta de los fenómenos psicológicos mediante tests.

Como respuesta a las restricciones impuestas por el modelo general de medida desarrollado en la física, la psicometría avanzó en el debate sobre la aproximación indirecta a los fenómenos psicológicos desarrollando sus propias teorías y modelos de medida. A pesar de que otros autores utilizan otros términos (por ejemplo, podéis ver Fraser, 1980, Swistak, 1990 o Hand 1996), en este texto seguiremos a Michell (1986) y los denominaremos *modelo representacional, operacional y clásico*. A continuación, caracterizaremos estos modelos y analizaremos brevemente las diferencias en su planteamiento del proceso de medida para hacerlo extensivo a los fenómenos psicológicos. Siendo un debate todavía abierto, y susceptible a diferentes interpretaciones según los diferentes autores, optamos por la exposición de Michell en cuanto que plantea, desde una visión crítica, las incógnitas más importantes que la psicometría debe resolver en su construcción como disciplina científica[6].

En primer lugar, el *modelo representacional* fue desarrollado a partir de los trabajos de Stevens (1946) y Suppes (1951), y define la medida como el proceso de asignación de números a objetos a partir de unas reglas, de modo que reflejen relaciones empíricas entre los objetos. Estas relaciones, de manera comparable al modelo de Campbell, quedarían representadas por las propiedades de los números, pero no se centrarían exclusivamente en la aditividad. De hecho, Stevens abrió la puerta a otros tipos de relaciones para llevar a cabo la medida, aceptando por ejemplo la equivalencia o el orden. Medir no sería únicamente cuantifi-

6. El lector interesado puede encontrar una discusión mucho más amplia de la que podemos plantear en este texto en Michell (1999).

car, sino la representación numérica de los hechos empíricos en sentido amplio, y daría como resultado cuatro escalas clásicas para la psicología según el tipo de relación representada: escala nominal, ordinal, de intervalo y de razón.

Tal y como hace el modelo desarrollado en la física, el modelo representacional asume la correspondencia entre las propiedades de los números –las puntuaciones obtenidas mediante tests, en este caso– y las relaciones de los objetos que representan. En cambio, traslada el foco desde la demostración empírica de estas relaciones al establecimiento de un conjunto de restricciones sobre el tipo de operaciones estadísticas permitidas en cada escala que garantice la invariancia de las relaciones empíricas representadas. Es el debate sobre la estadística permitida, que ha generado una gran controversia entre los partidarios del modelo representacional y el operacional[7].

En segundo lugar, el *modelo operacional* se basa en las contribuciones de Bridgman (1927) y Dingle (1950), a partir de las cuales se propuso la definición de cualquier concepto mediante las operaciones necesarias para medirlo. Así, de acuerdo con este modelo, la medida no sería otra cosa que el conjunto de operaciones necesarias para definir un concepto que, en última instancia, acaban produciendo números. Esta definición de la medida tiene puntos en común con la del modelo representacional, dado que asignar números –las puntuaciones obtenidas mediante tests– de acuerdo con unas reglas es un tipo particular de las operaciones posibles, y los números resultantes no son otra cosa que el producto de las operaciones llevadas a cabo. En cambio, una importante diferencia entre estos dos modelos de medida se encuentra en el requisito según el cual los números representan o no un sistema de relaciones empíricas.

Para el modelo representacional, estas relaciones empíricas son previas a la medida y tienen una existencia independiente de las operaciones llevadas a cabo para producir los números. El modelo operacional, en cambio, limita el dominio de las operaciones a aquellas que de una manera consistente producen números y en ningún caso está interesado en la existencia de una realidad empírica que apoye estos números. Esta diferencia en el enfoque de la medida tiene además importantes implicaciones en cuanto al debate sobre la estadística permitida, dado que los partidarios del modelo operacional no ven la necesidad de

7. En relación con el debate en torno a la estadística permitida, podéis ver la interesante discusión de Gaito (1980) y Velleman y Wilkinson (1993).

establecer ninguna restricción a las operaciones estadísticas. No siendo el objetivo la búsqueda de relaciones empíricas comparables a las relaciones entre los números, sino los números en sí mismos, ninguna restricción puede ser aplicable a unas puntuaciones que, en palabras de Lord (1953), no tienen conciencia de dónde provienen.

Finalmente, el *modelo clásico* se propondría resolver esta discusión planteando la naturaleza cuantitativa de los fenómenos psicológicos como condición para la medida. Partiendo de los trabajos de Rozeboom (1966) y Jones (1971), y arraigado en el desarrollo de las teorías cuantitativas de la psicología de las décadas anteriores, este modelo define la medida como la determinación de la cantidad –del cuánto– en la que un atributo esté presente en el objeto medido. Es decir, medir consiste en determinar cuántas unidades están presentes en el atributo observado, pero a diferencia de los modelos de Campbell y Stevens, no exige la existencia de una relación empírica entre los objetos. De hecho, el modelo clásico mide atributos, no objetos, y por lo tanto no contempla en su núcleo la demostración de la aditividad mediante la observación y la manipulación empírica de los objetos físicos. Es más, a diferencia de los modelos representacional y operacional, la medida no se considera la asignación de números a objetos de acuerdo con unas reglas ni el resultado de un tipo particular de operaciones que produce números, respectivamente, sino que se define como el proceso de descubrimiento de relaciones numéricas entre los valores observados en un atributo cuantitativo.

Siguiendo el ejemplo del peso de dos o más objetos físicos, sus medidas en una unidad determinada (por ejemplo, en kilos) proporcionan un sistema para relacionarlos matemáticamente en términos de sus magnitudes relativas, y permiten afirmar que uno de ellos pesa *n* veces más o menos que otro. Estas relaciones numéricas son, para el modelo clásico, tan válidas como las relaciones observables entre los objetos y por lo tanto comparten el mismo estatus como evidencias empíricas necesarias para el desarrollo científico. En este sentido, es muy importante señalar el papel de la teoría sustantiva, puesto que es la base sobre la que se fundamentan tanto la hipótesis sobre la naturaleza cuantitativa de los fenómenos psicológicos objeto de medida como el proceso de determinación de las relaciones numéricas a partir de las medidas obtenidas. Finalmente, el modelo clásico no contempla las diferentes escalas de medida aceptadas por los modelos representacional y operacional. De esta manera, las escalas nomi-

nales y ordinales quedan fuera del dominio de la medida –que, recordemos, es cuantitativa por definición– y por lo tanto no es necesaria ninguna prescripción en torno al debate sobre la estadística permitida.

Como ya hemos señalado, los modelos de medida psicométrica se encuentran en la base de las diferentes teorías de los tests y durante las últimas décadas han proporcionado varias alternativas para la aproximación indirecta a la medida de los fenómenos psicológicos que se propone la psicometría. Así, estos tres modelos de medida –decíamos, representacional, operacional y clásico– se corresponden con la *teoría representacional*, la *teoría clásica de los tests* y los *modelos de variable latente*, respectivamente. Abordar todas estas teorías alternativas está fuera de los objetivos de un texto introductorio como este, aunque el lector interesado puede encontrar un análisis muy detallado, por ejemplo, en Borsboom (2005). En cambio, optaremos por restringir el tratamiento de las teorías de los tests a aquella que goza del apoyo mayoritario en la práctica actual de la psicometría. Hablamos de la teoría clásica de los tests (TCT), que es además la base para el desarrollo de los capítulos posteriores de este texto. Es importante resaltar, para evitar confusiones, que la TCT no se denomina así por el modelo de medida psicométrica que utiliza. De hecho, su definición de medida encaja con el modelo operacional y, en cambio, adopta este nombre por oposición a otros enfoques más modernos, como el de la teoría de respuesta al ítem (TRI). Finalmente, introduciremos muy brevemente algunas de las innovaciones que, de acuerdo con el modelo clásico de medida propio de los modelos de variable latente, la TRI ha desarrollado para tratar de resolver algunas de las limitaciones de la TCT.

2.3. Teoría clásica de los tests

La TCT es la teoría de los tests más extendida actualmente en la práctica de la psicometría y se basa en el *modelo lineal clásico* propuesto por Spearman (para una discusión histórica, podéis ver Traub, 1997), sistematizado por Gulliksen (1950) y reformulado posteriormente por Lord y Novick (1968), que articula el proceso de medida definiendo tres conceptos fundamentales: la puntuación

verdadera, la puntuación empírica y el error de medida[8]. Para introducirlo, en este texto seguiremos la exposición de Muñiz (1996 y 2003).

Así, partiendo del modelo operacional de medida psicométrica, esta teoría de los tests no está interesada en el sistema de relaciones empíricas, sino que centra su atención en el análisis de las puntuaciones obtenidas para valorar los errores cometidos en el proceso de medida indirecta de los fenómenos psicológicos. Es la llamada *puntuación empírica* (*X*), que, de acuerdo con esta teoría, respondería a una relación lineal de dos componentes fundamentales:

$$X = V + e$$

Por un lado, la *puntuación verdadera* (*V*), que sería el resultado ideal o deseado, en el que el proceso de medida mediante tests habría sido llevado a cabo libre de cualquier tipo de error. Por otro lado, el *error de medida* (*e*), que sería responsable de la discrepancia entre la puntuación verdadera que se pretende conseguir y la puntuación empírica obtenida como resultado de la administración del test.

Como cualquier tipo de medida científica, la medida indirecta de los fenómenos psicológicos se encuentra sujeta a estas variaciones o desviaciones no deseadas, y el modelo lineal clásico propone un conjunto de supuestos que permiten hacer una estimación de las puntuaciones verdaderas a partir de las puntuaciones empíricas obtenidas, y una definición de tests paralelos. Así:

– **Primer supuesto**: $\boldsymbol{V = E(X)}$**.** La puntuación verdadera (*V*) se define matemáticamente como la esperanza matemática de la puntuación empírica (*X*). Es decir, el primer supuesto del modelo lineal clásico asume que, en caso de que fuera posible la administración de un test un número infinito de veces, la media de las puntuaciones empíricas obtenidas nos daría como resultado la puntuación verdadera del sujeto en el test. La puntuación empírica, por lo tanto, no es un sustituto de la puntuación verdadera, sino la mejor aproximación disponible en un proceso de medida que, de manera explícita, es reconocido como no libre de error.

8. El lector interesado puede encontrar explicaciones más detalladas en Borsboom (2005), en Crocker y Algina (2006), o en de Gruijter y van der Kamp (2008).

- **Segundo supuesto: $\rho(\upsilon, e) = 0$.** No existe correlación entre las puntuaciones verdaderas de los sujetos (υ) en un test y sus respectivos errores de medida (e). Es decir, de acuerdo con el segundo supuesto no se espera que el tamaño de los errores cometidos esté sistemáticamente asociado al tamaño de las puntuaciones verdaderas.
- **Tercer supuesto: $\rho(e_j, e_k) = 0$.** Si disponemos de dos tests diferentes (j y k), no existe correlación entre los errores de medida cometidos con cada uno de ellos (e_j y e_k, respectivamente). Es decir, de acuerdo con el tercer supuesto, los errores de medida de los diferentes tests son aleatorios en cada ocasión y, por lo tanto, no se espera que exista ninguna relación entre ellos.
- **Definición de tests paralelos: $V_j = V_k$ y $\sigma^2(e_j) = \sigma^2(e_k)$.** Finalmente, dos tests (j y k) son paralelos siempre que sus puntuaciones verdaderas (V_j y V_k) y sus varianzas de los errores de medida ($\sigma^2(e_j)$ y $\sigma^2(e_k)$, respectivamente) sean idénticas.

De este modelo, con sus supuestos y la definición de tests paralelos, se deriva un conjunto de deducciones inmediatas importantes que forman la base para el capítulo dedicado a la fiabilidad en este texto. Así:

- $e = X - V$. De acuerdo con la formulación inicial del modelo lineal clásico, el error de medida sería la diferencia entre la puntuación empírica y la puntuación verdadera.
- $E(e) = 0$. La esperanza matemática de los errores de medida es cero, por lo que si fuera posible administrar un test un número infinito de veces, estos errores aleatorios o no sesgados se compensarían o anularían entre ellos.
- $\mu_x = \mu_\upsilon$. La media de las puntuaciones empíricas (μ_x) es igual a la media de las puntuaciones verdaderas (μ_υ).
- $\text{cov}(V, e) = 0$. De acuerdo con el segundo supuesto del modelo, las puntuaciones verdaderas no covarían con los errores de medida.
- $\text{cov}(X, V) = \text{var}(V)$. La covarianza entre las puntuaciones empíricas y verdaderas es igual a la varianza de las puntuaciones verdaderas.
- $\text{cov}(X_j, X_k) = \text{cov}(V_j, V_k)$. La covarianza entre las puntuaciones empíricas de dos tests (X_j y X_k) es igual a la covarianza de sus puntuaciones verdaderas (V_j y V_k).

- **var** (X) = **var** (V) + **var** (e). La varianza de las puntuaciones empíricas es el producto de la suma de la varianza de las puntuaciones verdaderas y la de los errores de medida.
- $\rho(X, e) = \sigma_e / \sigma_x$. La correlación entre las puntuaciones empíricas y los errores de medida es el resultado de la división de la desviación típica de los errores entre la de las puntuaciones empíricas.
- $\mu_1 = \mu_2 = ... = \mu_k$. Para *k* tests paralelos, las medias son idénticas.
- $\sigma^2(X_1) = \sigma^2(X_2) = ... = \sigma^2(X_k)$. Del mismo modo, las varianzas de *k* tests paralelos son idénticas.
- $\rho(X_1, X_2) = \rho(X_1, X_3) = ... = \rho(X_j, X_k)$. Finalmente, las correlaciones entre *k* tests paralelos son también idénticas.

Teniendo en cuenta que ni las puntuaciones verdaderas ni los errores de medida son observables directamente a partir de las puntuaciones empíricas, ninguno de los supuestos ni las ocho primeras deducciones inmediatas son demostrables[9].

Como ya hemos discutido anteriormente, la TCT resulta una aproximación útil para la estimación de las puntuaciones verdaderas a partir de las puntuaciones empíricas, pero no está exenta de *algunas limitaciones*, que no son fáciles de resolver desde el modelo lineal clásico. Tal y como discutieron Lord y Novick (1968) en su reformulación, esta teoría tiene dos dependencias importantes en relación con los instrumentos y las muestras utilizadas. Por un lado, una dependencia de las puntuaciones empíricas obtenidas respecto a los instrumentos, de manera que, por ejemplo, dos tests de inteligencia independientes utilizan escalas diferentes y, por lo tanto, las puntuaciones que se obtienen no resultan comparables directamente. Por otro lado, una dependencia de las propiedades psicométricas de los tests respecto a las muestras utilizadas para desarrollarlos, de manera que, por ejemplo, la dificultad de los ítems que conforman un test depende de las características de los individuos a quienes se administra el test.

Tratando de resolver estos problemas, la TRI se propone cambiar el foco desde el tratamiento del test entero al tratamiento individual de los ítems. Para hacerlo, esta teoría de los tests plantea la existencia de una relación entre las

9. El lector interesado puede encontrar el desarrollo matemático que sustenta estas deducciones en Muñiz (2003).

puntuaciones en la variable latente –recordemos, de acuerdo con el modelo clásico de medida psicométrica– y la probabilidad de acertar cada ítem introduciendo diferentes funciones matemáticas para modelar adecuadamente esta relación. Así, una vez seleccionada la función más adecuada se construyen lo que se denominan las *curvas características de los ítems*, una modelización de las respuestas que puede tener en cuenta diferentes parámetros, como la capacidad de discriminación de los ítems, su dificultad o la probabilidad de que sean acertados al azar.

No es este el lugar para llevar a cabo una discusión en profundidad, pero sí para hacer énfasis en las implicaciones que tienen los diferentes modelos de medida psicométrica, a partir de los cuales la TCT y la TRI se proponen la medida indirecta mediante el uso de tests. Por un lado, tal y como hace la TCT, definiendo las puntuaciones empíricas como la suma de las puntuaciones verdaderas y el error de medida y, por otro, tal y como hace la TRI, estableciendo diferentes funciones matemáticas según el modelo de relación entre la variable latente y la capacidad de discriminación de los ítems[10].

2.4. El proceso de inferencia psicométrica

Para concluir con sus fundamentos, recapitularemos el proceso que la psicometría sigue para establecer sus inferencias sobre los fenómenos psicológicos no observables a partir de las puntuaciones obtenidas mediante tests. Antes de abordar sus especificidades, es importante tener presente que la medida objetiva y estandarizada de una muestra de comportamientos se ajusta, de manera general, al procedimiento establecido por el *método científico*. Lo podríamos resumir así:

1) Formular una pregunta de investigación o una hipótesis relevante.

10. El lector interesado puede encontrar una comparación asequible de estas dos teorías de los tests en Muñiz (2010), así como desarrollos más detallados de la TRI en Muñiz y Hambleton (1997), Maydeu-Olivares y McArdle (2005), Borsboom (2005) o Martínez Arias, Hernández Lloreda y Hernández Lloreda (2006). Una referencia clásica sobre los fundamentos de la TRI es la de Hambleton, Swaminathan y Rogers (1991).

2) Especificar y definir adecuadamente todas las variables involucradas.
3) Desarrollar o elegir los instrumentos y procedimientos necesarios para llevar a cabo las medidas.
4) Evaluar el funcionamiento de los instrumentos y procedimientos para obtener las garantías suficientes sobre la calidad del proceso de medida.
5) Recoger las evidencias necesarias que permitan responder a los objetivos de la investigación.
6) Resumir y, siempre que sea posible, tratar estadísticamente los datos obtenidos para determinar hasta qué punto los resultados son significativos y, por lo tanto, no son producto del azar.

De acuerdo con este procedimiento general, la psicometría ofrece un conjunto de teorías, métodos y técnicas vinculadas al desarrollo y la administración de tests que dan el soporte necesario a los puntos 2, 3 y 4 del método científico cuando se utiliza en el contexto de la medida indirecta de los fenómenos psicológicos no observables. Y lo hace, como hemos señalado anteriormente, partiendo de una definición precisa de los fenómenos objeto de medida y de una selección de los comportamientos implicados que serán empleados como evidencias observables. Aun así, es importante hacer énfasis en algunas dificultades específicas que, como disciplina científica, la psicometría debe afrontar para garantizar la confianza en las inferencias establecidas sobre los fenómenos psicológicos no observables a partir de las puntuaciones obtenidas mediante tests. Todas ellas, en íntima relación con el reto esencial de la psicometría, son consecuencia directa de la dificultad añadida que supone la imposibilidad de observar y manipular directamente los fenómenos psicológicos objeto de interés. En este sentido, entre los *retos específicos* podemos señalar algunos de los más importantes:

– *La medida indirecta de los fenómenos psicológicos no es unívoca*. Como hemos ido viendo, no existe una única manera de definir el proceso de medida, así como no existe tampoco una única manera de entender los propios fenómenos objeto de medida. Son muchas las decisiones que los profesionales interesados en el desarrollo y la administración de tests deben tomar durante este proceso y, por lo tanto, son muchas las posibles soluciones a los problemas que surgen en la aproximación indirecta a la medida de los fenómenos psicológicos. Entre estas decisiones destacan las relacionadas con la elección

de lo que es relevante medir, la selección y el muestreo de los comportamientos observables vinculados y la definición de las características que los tests deben tener para suscitarlos y medirlos adecuadamente.

– *La teoría desempeña un papel fundamental en la medida psicométrica*. Todas las decisiones que toman los profesionales interesados en el desarrollo y la administración de tests han de estar fundamentadas, como hemos señalado anteriormente, por dos tipos de teorías. Por un lado, por una teoría sustantiva muy establecida en torno a los fenómenos psicológicos y los comportamientos de las personas, que servirá de contexto de referencia para el proceso de medida. Por otro, una teoría de los tests que permitirá adoptar el enfoque más adecuado para llevar a cabo la medida y establecer la relación entre las puntuaciones obtenidas y los fenómenos psicológicos no observables, objeto de interés. Si los tests tienen sentido y sirven como instrumentos de medida en el campo de la psicología es, precisamente, gracias al apoyo que estos dos tipos de teorías proporcionan durante todo el proceso.
– *La selección y el muestreo de los comportamientos impone importantes limitaciones*. Un tercer reto importante para la medida indirecta que se propone la psicometría es la selección y el muestreo de los comportamientos empleados como evidencias observables. En consonancia con la teoría sustantiva, el desarrollo y la administración de tests parten de una definición que orienta hacia unos comportamientos vinculados y no a otros, que, a la vez, deben ser adecuadamente muestreados dada la imposibilidad material de suscitar y medir el dominio entero de comportamientos al que pertenecen. De hecho, un comportamiento determinado puede ser empleado como evidencia de fenómenos psicológicos diferentes en función de la perspectiva teórica con la que se fundamenta el proceso de construcción de los tests. Importantes limitaciones vinculadas al tiempo requerido y las condiciones de administración también han de ser adecuadamente resueltas para hacer operativos los tests desde un punto de vista práctico.
– *Las puntuaciones obtenidas requieren una interpretación adecuada*. Las puntuaciones obtenidas mediante el uso de tests no son, *per se*, informativas y han de ser siempre interpretadas para responder al propósito con el que los tests han sido desarrollados. Ya sea desde un punto de vista general, con relación al contenido y su propósito, ya sea con relación al sistema de referencia empleado –según la norma, es decir, respecto a la ejecución de un grupo de re-

ferencia, o referidos a un criterio, esto es, respecto al grado de adecuación o consecución de este criterio–, el uso de tests requiere unos conocimientos y unas destrezas que solo podemos encontrar en manos de profesionales cualificados y que siempre debe cumplir con los objetivos más generales que persigue la evaluación psicológica a la que sirve.

– *La medida de los fenómenos psicológicos no está libre de error*. El error es un componente inherente a cualquier proceso de medida y debe ser objeto de un tratamiento adecuado para conocerlo y minimizarlo. La medida indirecta añade una dificultad específica que, de manera sintética, encontramos en la imposibilidad de observar directamente tanto los fenómenos psicológicos objeto de interés como el error cometido durante el proceso de medida. Partiendo de las puntuaciones obtenidas, y gracias al apoyo de los modelos de medida desarrollados en las diferentes teorías de los tests, la psicometría dedica una buena parte de sus esfuerzos a conseguir la precisión necesaria que, desde un punto de vista científico, es también exigible a la medida indirecta de los fenómenos psicológicos.
– *La medida psicológica no se puede entender si no es en relación con otras medidas o acontecimientos observables*. Pese a la importancia de la teoría sustantiva en el desarrollo de los tests, no podemos obviar finalmente que la medida indirecta de los fenómenos psicológicos no tiene sentido solo en sí misma, sino que lo tiene en relación con otras medidas derivadas de la misma teoría u otros comportamientos observables que corroboren el proceso de medida mediante tests. Como hemos dicho anteriormente, la medida mediante tests debe proporcionar información relevante con el objetivo de hacer descripciones, clasificaciones, diagnósticos, explicaciones o predicciones que permitan orientar una acción o la toma de una decisión sobre el comportamiento de las personas, y por esa razón las puntuaciones obtenidas mediante tests requieren un apoyo externo a su propia formulación para demostrar su utilidad práctica.

Para enfrentarse a estos retos específicos, la medida indirecta de los fenómenos psicológicos se provee de unos *principios básicos* que garantizan la confianza en las inferencias establecidas a partir de las puntuaciones obtenidas mediante tests. Estos principios sirven para evaluar las propiedades psicométricas de los

ítems y de los tests en su conjunto, y entre estos principios básicos destacan los siguientes:

- *Fiabilidad*: la precisión con la que los tests llevan a cabo la medida.
- *Validez*: la confianza en que las medidas se corresponden realmente con lo que se proponen medir.

Los diferentes métodos y procedimientos vinculados a estos principios son la clave del éxito de los tests como instrumentos de evaluación psicológica y serán, por lo tanto, objeto de un tratamiento en detalle en capítulos posteriores.

3. Construcción y administración de tests

Presentados los fundamentos de la psicometría, en este último apartado nos haremos cargo del proceso de construcción y administración de tests. Para hacerlo, empezaremos desarrollando las diferentes fases implicadas en el diseño y construcción de un nuevo test. A continuación, y partiendo de las claves que nos proporciona el conocimiento de este proceso de construcción, discutiremos algunos criterios importantes para evaluar las características de los tests disponibles en la literatura y valorar su conveniencia en relación con los objetivos de la evaluación psicológica a la que deben servir. Finalmente, abordaremos los aspectos éticos y deontológicos vinculados al uso de tests en el contexto general de la práctica profesional de la psicología.

3.1. El proceso de construcción de tests

El desarrollo de instrumentos de medida es un proceso fundamental para cualquier disciplina científica. Y no lo es menos para la psicometría, en la que, como ya hemos discutido ampliamente, la imposibilidad de observar y manipular los fenómenos objeto de interés añade una importante complejidad. A continuación abordaremos *diez fases fundamentales* en las que, a modo de con-

clusión, podemos organizar el proceso de construcción de un nuevo test. Estas fases nos permiten dar una perspectiva de conjunto a esta introducción a la psicometría y elaborar una síntesis de los conceptos más importantes que hemos ido discutiendo[11].

A pesar de que no son necesariamente secuenciales, y en todo caso no tienen por qué seguir estrictamente este orden, estas diez fases representan algunas de las decisiones más importantes que los profesionales interesados en el desarrollo de nuevos tests deben tomar. Como veremos más adelante, conocer este proceso es fundamental no solo para garantizar la calidad en el diseño y la construcción de nuevos tests, sino porque nos permite desarrollar algunos criterios importantes para la evaluación de los tests disponibles en la literatura. Así:

1) Delimitación de la finalidad del test. El desarrollo de un nuevo test empieza con una determinación clara del propósito para el que se pretende recoger información relevante en el contexto de la evaluación psicológica. Tal y como hemos señalado, los tests pueden servir a multitud de finalidades, pero entre ellas podemos destacar algunas importantes, como son describir, clasificar, diagnosticar, explicar o hacer predicciones sobre el comportamiento de las personas. Un propósito bien delimitado es la primera condición que hay que cumplir para garantizar el éxito en la construcción del nuevo test y permitirá su introducción en el proceso general de evaluación psicológica.

2) Definición de los fenómenos psicológicos objeto de medida. De acuerdo con el proceso de inferencia psicométrica, el segundo paso para la construcción de un test consiste en la delimitación precisa de los fenómenos que se pretenden medir. Para hacerlo, recordémoslo, es necesaria una teoría sustantiva sólida y bien contrastada sobre los fenómenos psicológicos y los comportamientos de las personas que servirá de referencia en el proceso de la medida mediante el nuevo test. Este marco teórico es fundamental en el momento de planificar una representación adecuada de los fenómenos psicológicos y permite conocer en detalle otros instrumentos desarrollados previamente en la investigación psi-

11. El lector interesado puede encontrar explicaciones más detalladas en Muñiz (1996), Muñiz y Fonseca-Pedrero (2008), Murphy y Davidshofer (2005), Downing (2006), Chadha (2009) o Rust y Golombok (2009).

cológica. Evaluar sus limitaciones y sus puntos fuertes supone una buena guía para enfocar el trabajo necesario para dar soporte teórico al nuevo test.

3) Selección y muestreo de los comportamientos observables. Una vez establecido el fenómeno objeto de medida, la teoría sustantiva proporciona también el contexto necesario para elegir los comportamientos implicados que serán empleados como evidencias observables. Su representación adecuada es fundamental para no omitir ningún comportamiento relevante, así como para evitar incluir otros no directamente relacionados con el objeto de medida. En ocasiones resulta recomendable también llevar a cabo observaciones, entrevistas a informadores clave o grupos de discusión que proporcionen información complementaria sobre los comportamientos de interés a partir de las experiencias de los participantes. Como ya hemos discutido, esta no es una fase menos importante en cuanto a que es la base para el correcto desarrollo de los ítems que conformarán el nuevo test.

4) Especificación de las características del test. Delimitado el objetivo, definido el objeto de medida y seleccionadas las evidencias observables necesarias, el siguiente paso consiste en elegir las características del nuevo test para suscitar y medir adecuadamente los comportamientos de interés. En primer lugar, decidiremos si se trata de un test de habilidad o de potencia, en el que se evaluarán el acierto y el error en las respuestas para determinar la competencia, la aptitud o los conocimientos de los individuos, o de una prueba para evaluar motivaciones, preferencias, opiniones o actitudes. Es decir, si se trata de una prueba de ejecución máxima o de ejecución típica, como también son conocidas. De acuerdo con la clasificación de los tests que hemos hecho en esta introducción a la psicometría, es el momento también para decidir el formato del nuevo test, optando por una prueba de lápiz y papel, de manipulación o de medidas fisiológicas. Asimismo, se deben preparar los materiales necesarios y se determinará el tipo de administración más adecuada, eligiendo básicamente entre una prueba individual o de administración en grupo. Finalmente, también se decidirá el método más adecuado para interpretar las puntuaciones obtenidas, ya sea en base a la norma o en referencia a un criterio.

5) Desarrollo de los ítems que conformarán el test. Una vez especificadas las características generales del nuevo test, en esta fase se llevará a cabo el desarrollo de los elementos que contendrá. Esta no es una tarea sencilla y generalmente supone la colaboración de un grupo de expertos en el campo para

encontrar la mejor representación de las muestras de comportamiento seleccionadas. De hecho, no es poco habitual desarrollar muchos más ítems de los estrictamente necesarios para evaluar su comportamiento en el marco del test y seleccionar los más idóneos con relación al propósito del nuevo test. El objetivo final es disponer de tantos como sean necesarios para representar adecuadamente las diferentes dimensiones de los fenómenos psicológicos objeto de interés. Este es el momento también para decidir el formato que adoptarán las respuestas, que servirán para codificar los comportamientos de manera estructurada y de acuerdo con unas reglas claras.

6) Elección de una teoría de los tests. Cerrando el círculo establecido en el proceso de inferencia psicométrica, la siguiente fase en el proceso de construcción implica la elección de una teoría de los tests que, mediante un modelo de medida psicométrica determinado, servirá para relacionar los fenómenos psicológicos objeto de interés y las puntuaciones obtenidas mediante los ítems que conforman el nuevo test. Esta elección es capital teniendo en cuenta sus consecuencias en los métodos y técnicas empleadas posteriormente para evaluar las propiedades de los ítems y del test en su conjunto con el objetivo de garantizar la confianza en las inferencias establecidas sobre el comportamiento de las personas a partir de las puntuaciones obtenidas.

7) Realización de una prueba piloto. Una vez construido el test, incluyendo los ítems potenciales y el formato de las respuestas, es necesario redactar las instrucciones que lo acompañarán y definir las condiciones en las que será administrado. Una prueba piloto servirá para evaluar el grado de comprensión de estas instrucciones, la viabilidad de las condiciones para administrar el test, detectar posibles dificultades en el momento de registrar las respuestas y llevar a cabo un primer análisis de las propiedades, tanto de los ítems como del test en su conjunto, de acuerdo con los métodos y las técnicas indicadas por la teoría de los tests utilizada. Es el momento para valorar su comportamiento en el proceso de medida y, a partir de esta información, refinar el test modificando, descartando o añadiendo nuevos ítems. Pruebas adicionales pueden ser requeridas para evaluar correctamente las modificaciones introducidas antes de cerrar esta fase y proceder con el desarrollo del estudio de campo final.

8) Desarrollo del estudio de campo. Una vez establecido el test definitivo, el siguiente paso consiste en su administración a la población a la que se dirige. Para hacerlo, se selecciona la muestra de participantes necesarios, que, como ya hemos

señalado, puede ser probabilística o no probabilística en función de si queremos representar la población de referencia o evaluar a un grupo de personas que cumplen unas determinadas características. Con esta información se desarrollan las normas o baremos para permitir la interpretación de las puntuaciones en relación con la ejecución del grupo de referencia. En el caso de que se trate de un test referido a criterio, en lugar de normas o baremos se determinan los puntos de corte que permitirán distinguir los diferentes grados de adecuación o consecución del criterio. Asimismo, se profundizará en el trabajo sobre las propiedades psicométricas de los ítems y del test en su conjunto, atendiendo especialmente a la fiabilidad y validez de las medidas obtenidas. Esta es una parte fundamental del proceso de construcción de tests y es la que en última instancia garantiza que la medida psicométrica cumple con todas las exigencias científicas.

9) Elaboración del manual del test. Con toda esta información se desarrolla la documentación que acompañará al nuevo test, donde se ha de incluir información relevante relativa a las diferentes fases involucradas en su construcción: fundamentación teórica, finalidad y población a la que se dirige, instrucciones para la administración, información para la interpretación de las puntuaciones obtenidas y análisis de las propiedades psicométricas. Las primeras publicaciones en revistas científicas sirven para empezar a difundir toda esta información, que como ya hemos comentado no siempre acaba constituyendo un manual propiamente dicho.

10) Revisión y mejora del test. Con la publicación del test, ya sea licenciándolo o difundiéndolo libremente, el test se pone a disposición de la comunidad científica para obtener nuevas evidencias, que, mediante el trabajo independiente de diferentes investigadores, servirán para mejorar el conocimiento sobre su funcionamiento y sus propiedades psicométricas, así como para adaptarlo a otros entornos socioculturales o a otras poblaciones diferentes para las que ha sido desarrollado el nuevo test. Cambios en los fenómenos psicológicos objeto de medida, nuevos avances en la aproximación teórica a estos fenómenos o la adaptación del test a nuevas condiciones de administración o poblaciones de interés son algunas de las razones que justifican un trabajo de revisión y actualización prácticamente indefinida para refinar su funcionamiento y valorar adecuadamente su utilidad al servicio de los objetivos de la evaluación psicológica.

3.2. Criterios para la valoración de tests

Construir un test nuevo no es la práctica más habitual en el ejercicio profesional de la evaluación psicológica. De hecho, como hemos visto, es un trabajo laborioso que requiere un importante esfuerzo por parte de varios profesionales durante un periodo de tiempo prolongado para planificar, ejecutar y analizar los datos recogidos en uno o más estudios de campo. Por ello la mayoría de las situaciones en las que se utilizan los tests no comienzan necesariamente con la creación de uno original. En cambio, parten de la búsqueda y valoración de algún test existente en la literatura que encaje con los objetivos de la evaluación psicológica. Para simplificar este proceso, organizaciones vinculadas al ejercicio profesional de la psicología como la American Psychological Association (http://www.apa.org) o el Consejo General de Colegios Oficiales de Psicólogos (http://www.cop.es), sugieren algunas *fuentes para la búsqueda de tests*:

- *Publicaciones periódicas*. Con formato libro o enciclopedia, varias publicaciones se dedican a recopilar los tests comerciales disponibles con diferente nivel de detalle. Por ejemplo, el Buros Center for Testing de la Universidad de Nebraska edita *Tests in Print* (TIP, http://buros.org/tests-print) y *Mental Measurements Yearbook* (MMY, http://buros.org/mental-measurements-yearbook). TIP es una de las recopilaciones en lengua inglesa más exhaustivas que proporciona información relevante sobre los objetivos, las poblaciones a las que se dirigen, los autores y editoriales que los publican y el precio de más de 3.000 tests. Por su parte, MMY va más allá de la estricta recopilación y dedica sus esfuerzos a proveer diferentes evaluaciones sobre la calidad de los tests. Otra referencia internacional importante es Pro-Ed (http://www.proedinc.com), que edita también con carácter periódico sus *Test Publisher*, con el objetivo de recopilar los tests comerciales disponibles, y *Test Critiques Publisher*, que funciona como un recurso complementario al anterior y proporciona también información sobre su calidad.
- *Bases de datos*. Por otro lado, en la Red podemos encontrar algunos directorios abiertos que proporcionan información sobre los tests indexados, tanto comerciales como no comerciales, y facilitan su acceso. Entre otros, podemos destacar la *Test Collection* (http://www.ets.org/test_link), del Educatio-

nal Testing Service, que incluye más de 25.000 tests. En este sentido, otra fuente interesante es *Test Reviews Online* (http://buros.unl.edu/buros/jsp/search.jsp), una base de datos desarrollada por el Buros Center for Testing, que permite la búsqueda electrónica entre los tests incorporados al MMY y que ofrece acceso comercial individual a sus evaluaciones.

– *Catálogos de las editoriales*. De especial interés para la búsqueda de tests en lengua no inglesa, y en especial en castellano, diferentes editoriales españolas publican los catálogos de tests que licencian y proveen de la información necesaria para conocer sus características. Entre estos, podemos destacar los catálogos de TEA Ediciones (http://www.teaediciones.com), del Grupo Albor-Cohs (http://www.psicologia365.com), del Instituto de Orientación Psicológica EOS (http://www.eos.es) y de Pearson (Pearson Clinical & Talent Assessment, http://www.pearsonpsychcorp.es).
– *Búsqueda de tests no comerciales*. Finalmente, debemos tener presente que no siempre los tests comerciales disponibles en el mercado se ajustan a las necesidades de la evaluación. En este caso, en la búsqueda de tests no comerciales para utilizarlos en la evaluación psicológica es conveniente consultar las bases de datos que incluyen las publicaciones científicas más relevantes en el ámbito de conocimiento de la psicología. Más allá de las recomendaciones habituales, podemos destacar *PsycTESTS* (http://www.apa.org/pubs/databases/psyctests), una base de datos especializada en tests no comerciales desarrollada por la American Psychological Association y actualizada con una periodicidad mensual, en la que se recopilan tests obtenidos de los propios autores, de publicaciones en revistas científicas, libros o tesis.

Una vez encontrado uno o más tests potencialmente interesantes, es necesario evaluar sus características para *valorar la conveniencia de su uso* en relación con los objetivos de la evaluación psicológica. Para hacerlo, podemos volver al proceso de construcción de tests expuesto en el apartado anterior y ofrecer algunos criterios a partir de las preguntas más importantes que los profesionales interesados en su uso deben formularse en relación con los objetivos del test, sus propiedades psicométricas, el proceso de administración, su estatus y, finalmente, la adecuación específica al proceso de evaluación:

- *En relación con los objetivos del test*. Entre otras cuestiones, ¿está claramente definido el objetivo del test? ¿Dispone del soporte teórico necesario? ¿Disponemos de las evidencias científicas necesarias para utilizarlo como instrumento de medida? ¿Se actualiza regularmente o al menos es objeto de investigación para valorar sus propiedades psicométricas y su utilidad según los objetivos que se propone?
- *En relación con sus propiedades psicométricas*. De acuerdo con el proceso de construcción, ¿se proporciona la información necesaria sobre el análisis de los ítems que componen el test? ¿Qué evidencias se proporcionan sobre la fiabilidad y la validez de las medidas obtenidas mediante el test? ¿Qué tipos de muestras, y de qué tamaños, se han utilizado para recoger estas evidencias?
- *En relación con el proceso de administración*. ¿Dispone el test de un manual adecuado con la información y los materiales necesarios para su uso? ¿Tiene instrucciones claras y/o plantillas para la corrección? Si fuera necesario, ¿este manual dispone de normas o baremos para la interpretación de las puntuaciones obtenidas?
- *En relación con su estatus*. ¿Se trata de un test comercial? Y en caso de serlo, ¿qué coste tienen el manual, las copias del test, las hojas de respuesta y la plantilla de corrección? ¿Es necesario algún tipo de calificación o acreditación para administrarlo? ¿Ha sido evaluado por alguna entidad u organismo independiente?
- *En relación con la adecuación al proceso de evaluación*. Finalmente, ¿es pertinente de acuerdo con los objetivos de evaluación? ¿Es su marco teórico congruente con estos objetivos? ¿Se dirige específicamente a la población a la que se quiere administrar? ¿Está adaptado o dispone de una versión adaptada al contexto cultural en el que quiere ser utilizado? ¿Se ajusta a las condiciones de administración previstas en el proceso de evaluación, por ejemplo, teniendo en cuenta el medio de administración, los materiales necesarios, el formato de respuesta a los ítems o el tiempo requerido?

Evaluar un test antes de utilizarlo es más que una práctica conveniente en el ejercicio profesional de la psicología y, de hecho, supone un importante reto al que las propias organizaciones profesionales han dedicado diferentes esfuerzos para sistematizar su abordaje. Sin detenernos en sus especificidades, y teniendo

en cuenta el diferente nivel de complejidad en el que han sido desarrolladas, estas iniciativas se basan también en el análisis de las decisiones tomadas en el proceso de diseño y construcción de los tests, y proporcionan un marco de evaluación de la calidad de los tests disponibles en la literatura muy útil para la valoración de su conveniencia con relación a los objetivos de la evaluación psicológica.

Así, en el contexto internacional, debemos destacar la iniciativa de la American Educational Research Association, la American Psychological Association y el National Council on Measurement in Education (1999), que desde 1985 publican conjuntamente sus *The standards for educational and psychological testing*. Por otro lado, la European Federation of Psychologists' Associations (2012) trabaja en el desarrollo del *EFPA Review model for the description and evaluation of psychological and educational tests* con la intención de armonizar los criterios y procesos de evaluación de la calidad de los tests a escala europea. Finalmente, en el contexto estatal, debemos hacer mención a la contribución del Consejo General de Colegios Oficiales de Psicólogos para adaptar el modelo del EFPA en "*Un modelo para evaluar la calidad de los tests utilizados en España*" (Prieto y Muñiz, 2000). Este modelo ya ha sido utilizado con éxito para evaluar la calidad de diez de los tests más utilizados por los psicólogos españoles (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez y Peña-Suárez, 2011).

3.3. Aspectos éticos y deontológicos en el uso de tests

Como hemos discutido ampliamente a lo largo de esta introducción a la psicometría, el uso de tests como instrumentos de evaluación psicológica es una de las prácticas habituales en el ejercicio de la psicología. Sea al servicio de la intervención en los diferentes campos profesionales, sea en el centro del desarrollo de los campos académicos y científicos, los tests son unos instrumentos indispensables para la medida objetiva y estandarizada de los fenómenos psicológicos. Y como tales se encuentran sujetos a una reflexión sobre las *consecuencias éticas y morales que implica su uso*. Es decir, en cuanto que instrumentos, pueden ser utilizados o no de manera adecuada y de acuerdo con unas finalidades u otras. No es este el lugar para hacer una discusión filosófica en profundidad sobre lo que significa actuar bien o hacer las cosas correctamente, pero sí para con-

cluir esta introducción reflexionando sobre algunas cuestiones importantes que tienen que ver con un uso de los tests responsable, justo y respetuoso con los derechos y la dignidad de las personas.

Para hacerlo, repasaremos algunos de los códigos de conducta más importantes desarrollados en el campo de la psicología que, en último término, presiden el correcto desarrollo de la actividad profesional, docente e investigadora. Y analizaremos el modo como abordan los aspectos éticos y morales vinculados al uso de tests en el contexto general de la evaluación psicológica. Estos códigos son la respuesta específica que la psicología ha dado a un debate complejo que, como otras disciplinas, ha afrontado a partir de la Segunda Guerra Mundial. Partiendo del Código de Nuremberg (1947), la Declaración de Helsinki (1964), el Informe Belmont (1979) y las directrices propuestas por el Council for International Organisations of Medical Sciences (1982), la consideración de las implicaciones éticas y morales del trabajo con humanos ha formado y forma parte de la práctica cotidiana de profesionales e investigadores (Israel y Hay, 2006). En el ámbito concreto de la psicología, son muchos los códigos que las organizaciones profesionales de los diferentes países han desarrollado, pero en este texto nos referiremos a los trabajos de la American Psychological Association, la European Federation of Psychologists' Associations y, en el caso español, el Consejo General de Colegios Oficiales de Psicólogos. Finalmente, comentaremos unas directrices internacionales que, en referencia al caso específico del uso de los tests, ha desarrollado la International Test Commission.

Como una de las referencias más importantes a nivel internacional, la American Psychological Association ha contribuido al debate sobre las implicaciones éticas y morales en el campo de la psicología declarando uno de los primeros códigos deontológicos. A partir de los años cincuenta, y en sucesivas revisiones hasta la actualidad, esta asociación promueve lo que denomina *Ethical principles of psychologists and code of conduct* (American Psychological Association, 2010), donde establece un preámbulo, cinco principios generales y diez estándares que establecen los límites y las pautas adecuadas para una práctica profesional responsable. El preámbulo y los cinco principios se presentan como un posicionamiento ético y moral, es decir, un compromiso con los ideales de la psicología y no tanto como unas normas de obligado cumplimiento. En este sentido, el código reclama la aspiración por la beneficencia y no maleficencia, la fidelidad y

responsabilidad, la integridad, la justicia y el respeto por los derechos y la dignidad de las personas.

En otras palabras, los psicólogos deben hacer todo lo posible para buscar y preservar el bienestar de las personas, evitando producir cualquier mal y velando por los derechos de aquellas personas con las que interactúan en los diferentes ámbitos de actividad profesional. Asimismo, deben desarrollar relaciones de confianza, responsabilizándose de su comportamiento y tratando de resolver activamente los conflictos de interés que puedan presentarse en el ejercicio profesional. Los psicólogos deben, además, actuar honestamente y con rigor, evitando en todo momento el engaño o el equívoco. Y tienen que hacerlo también con equidad y justicia, dedicando un especial cuidado a asegurar que ningún sesgo ni límite en su competencia o en su experiencia profesional contribuya o apoye el desarrollo de prácticas injustas. Finalmente, los psicólogos han de velar por los derechos de las personas a la privacidad, confidencialidad y dignidad, respetando las diferencias culturales e individuales, y evitando contribuir o dar apoyo a cualquier tipo de prejuicio en sus prácticas profesionales.

Por su parte, los estándares se presentan como un conjunto de normas de obligado cumplimiento para los miembros de la asociación y los somete a las decisiones o sanciones que su comité ético pueda decidir en caso de violación del código en el ejercicio profesional. Estos estándares están descritos de manera general y con diferente nivel de articulación, con el objetivo de que puedan ser aplicados de manera transversal a los diferentes contextos de actividad profesional, docente o investigadora. Los estándares remiten a diez aspectos fundamentales, como son la resolución de los conflictos éticos, la competencia profesional, las relaciones humanas, la privacidad y la confidencialidad, la publicidad y las contribuciones públicas, el mantenimiento de la documentación profesional y los honorarios, la formación, la investigación y la publicación de los resultados, la evaluación y la terapia[12].

En este sentido, y de acuerdo con los objetivos de esta discusión sobre los aspectos éticos y deontológicos vinculados al uso de tests, el estándar relativo a la evaluación describe con detalle el marco general en el que esta debe ser llevada a

12. El lector interesado puede encontrar en Campbell, Vasquez, Behne y Kinscherff (2010) un comentario detallado de estos estándares, así como unos dilemas a modo de ejemplo de su aplicación a la práctica.

cabo, y plantea la necesidad de basar cualquier juicio o conclusión en las evidencias obtenidas mediante el uso apropiado de los instrumentos de evaluación. Además, hace referencia explícita al uso de tests y plantea la necesidad de utilizar tests válidos y fiables, la obligación de poner a disposición de las personas sus respuestas y su interpretación de manera confidencial, el desarrollo de nuevos tests siempre de acuerdo con los estándares de calidad de la psicometría, la desestimación de tests obsoletos o no adecuados para la población objetivo y el cumplimiento de las condiciones y licencias de uso de los tests desarrollados por terceros.

Por su parte, la European Federation of Psychologists' Associations (2005) ha propuesto un *Meta-Code of Ethics*, que sirve como marco de referencia para el contexto europeo y pretende homogeneizar el tratamiento de las cuestiones éticas y deontológicas de las diferentes organizaciones profesionales de los Estados miembros. Para hacerlo, este código contiene un preámbulo general en el que establece su obligación de velar por el desarrollo profesional de los psicólogos europeos en todas estas cuestiones y propone cuatro principios fundamentales que, articulados en un conjunto de recomendaciones, deben regir los diferentes códigos, procedimientos y comités éticos nacionales.

Así, el código europeo plantea la defensa y el respeto por los derechos y la dignidad de las personas, el reconocimiento de la importancia de la competencia de los profesionales que ejercen la psicología, la responsabilidad frente a las personas, las comunidades y la sociedad en general y la promoción de la integridad en el desarrollo de la actividad profesional, docente e investigadora. Tanto los principios como las diferentes especificaciones se consideran interdependientes, por lo que se plantea abiertamente la necesidad de establecer un debate y diálogo profesional ante la complejidad de los conflictos éticos que se presentan en la práctica. El contenido de su articulado reclama la atención de los psicólogos hacia cuestiones importantes relativas al ejercicio profesional cotidiano comparables a las establecidas en el código de la American Psychological Association, pero a diferencia de este, no hace ninguna referencia explícita a los procesos de evaluación psicológica ni al uso de tests en este contexto.

Por otro lado, en relación con el ejercicio profesional en el campo de la psicología en España, el Consejo General de Colegios Oficiales de Psicólogos (2010) ha desarrollado su *Código deontológico* como la plasmación de unos derechos y deberes profesionales que, en última instancia, sirven de base para el juicio de la conducta de sus colegiados (Bermejo, 2009, para una discusión de las

modificaciones más recientes). Dispone de un título preliminar, a modo de preámbulo, que determina su alcance y enfatiza tanto el respeto fundamental por el marco normativo y jurídico, como la obligación de rechazar cualquier tipo de impedimento o limitación en el ejercicio profesional libre e independiente. Propone también unos principios generales que, a continuación, desarrolla en un conjunto de áreas fundamentales, y un anexo final recoge el reglamento de la comisión deontológica estatal encargada de velar por la correcta interpretación y aplicación del código. Así, entre los principios generales que deben regir la actividad profesional de los psicólogos en el Estado español, plantea:

1) La atención primordial al bienestar, la salud, la calidad de vida y la plenitud del desarrollo de las personas.
2) La protección de los derechos humanos y la responsabilidad en la fundamentación objetiva y científica de las intervenciones profesionales.
3) El no desarrollo o contribución a prácticas que atenten contra la libertad y la integridad física y psíquica de las personas.
4) La obligatoriedad de informar al menos a los organismos colegiales ante el conocimiento de cualquier violación de los derechos humanos o tratamiento degradante.
5) El respeto por los criterios morales y religiosos de las personas.
6) La no discriminación por cualquier diferencia o motivo.
7) La denegación de cualquier beneficio o provecho que se pueda extraer como consecuencia de las relaciones de poder o superioridad establecidas.
8) El uso de un lenguaje prudente y crítico ante etiquetas despreciativas o discriminatorias en sus informes.
9) El respeto por la actividad de los otros profesionales y la libre competencia.
10) La denuncia de prácticas ilegítimas o intrusivas.
11) La imparcialidad ante los posibles conflictos de interés entre los psicólogos o las instituciones en las que desarrollan su actividad y los intereses de las personas.

El articulado desarrolla estos principios en siete áreas fundamentales: la competencia profesional y la relación con otros profesionales, la intervención, la investigación y la docencia, la obtención y uso de la información, la publicidad, los honorarios y la remuneración, y las garantías procesales. Articuladas con di-

ferente nivel de detalle, estas áreas no contemplan específicamente la evaluación psicológica ni hacen ninguna referencia explícita al uso de tests. Estas cuestiones quedan relegadas a un tratamiento genérico en las áreas de la competencia profesional y de la obtención y uso de la información, de manera que se plantea la necesidad de disponer de los conocimientos necesarios para el uso de los métodos, los instrumentos, las técnicas y los procedimientos, siempre de acuerdo con las evidencias científicas necesarias que garanticen un uso adecuado. Con relación a la obtención y uso de la información, circunscribe el uso de estos métodos, instrumentos, técnicas y procedimientos de acuerdo con las condiciones de confidencialidad y secreto profesional, así como reconoce el derecho de las personas a conocer sus resultados.

Finalmente, más allá de los códigos deontológicos profesionales, debemos hacer mención al trabajo específico que la International Test Commission (2000) ha desarrollado en sus *International guidelines for test use*. Dada la disparidad en la normativa y el nivel de desarrollo de los códigos profesionales de los diferentes países, esta organización internacional plantea unas directrices específicas sobre el uso de los tests y su papel en el contexto de la evaluación psicológica que, entre otros, subscribe el Consejo General de Colegios Oficiales de Psicólogos. Tal y como señala en su introducción, la intención de estas directrices no es añadir otro conjunto de recomendaciones a las ya existentes, sino que persigue crear una estructura coherente bajo la que se puedan entender y aplicar los diferentes códigos y estándares nacionales que desarrollan los aspectos éticos y deontológicos vinculados al uso de tests. Para hacerlo, estructura su articulado a partir de un propósito general que establece un uso apropiado, profesional y ético de los tests, con respeto a los derechos de las personas, las razones por las que se utilizan y el contexto en el que se aplican.

Asimismo, como gran eje conductor de sus directrices, propone la consecución de este propósito general mediante el desarrollo y la adquisición de las competencias necesarias para llevar a cabo la administración de tests, la interpretación y comunicación adecuadas de los resultados y la resolución de las dificultades, los malentendidos y los conflictos que se puedan producir durante el proceso. De este modo, estas directrices establecen que los usuarios competentes de los tests deberían:

1) Responsabilizarse del uso ético de los tests:
- Actuando de manera profesional y ética.

- Asegurando que tienen los conocimientos y las habilidades necesarias.
- Haciéndose responsables del uso de los tests.
- Manteniendo de forma segura los materiales de los tests.
- Garantizando la confidencialidad de los resultados.

2) Comprometerse con las buenas prácticas en el uso de los tests:
- Estudiando la utilidad de los tests en los procesos de evaluación.
- Eligiendo tests bien fundamentados y apropiados para la situación.
- Atendiendo a las cuestiones relacionadas con la equidad.
- Preparando adecuadamente las condiciones de administración.
- Administrando correctamente los tests.
- Obteniendo las puntuaciones y analizándolas con exactitud.
- Interpretando los resultados correctamente.
- Comunicando los resultados de manera clara y exacta.
- Evaluando adecuadamente el funcionamiento y las propiedades de los tests.

Todas estas directrices están, a su vez, desarrolladas en un conjunto de recomendaciones concretas y representan un complejo retrato de las dificultades y los retos éticos y morales que plantea específicamente el uso de los tests como instrumentos de evaluación psicológica. Finalmente, el texto concluye con unos anexos, que desarrollan algunas indicaciones para el tratamiento específico de cuatro cuestiones importantes, como son el desarrollo de políticas y normativas sobre el uso de tests en las organizaciones y empresas, la redacción de contratos entre las diferentes partes involucradas en el uso de los tests, la administración de tests a personas con discapacidades y la traducción de las propias directrices a los diferentes idiomas por parte de las organizaciones estatales vinculadas al ejercicio profesional de la psicología.

Como hemos podido destacar, las cuestiones éticas y morales forman parte intrínseca del ejercicio profesional de la psicología. Así lo reflejan los diferentes códigos deontológicos que rigen la actividad de los psicólogos en Estados Unidos, Europa y España, mediante los cuales se reconoce el reto que supone afrontar de manera activa, responsable y comprometida la complejidad de las situaciones y los contextos en los que los profesionales intervienen de manera

cotidiana. Por esta razón es obligación de los psicólogos exigirse, y exigir a los otros profesionales con quienes colaboran, un comportamiento ejemplar ante los dilemas éticos y morales que la práctica pueda conllevar. Los tests, instrumentos imprescindibles para la evaluación y la intervención psicológicas, no se encuentran al margen de esta discusión a pesar de que, como hemos comentado, los códigos no siempre hacen un tratamiento específico de los procesos de evaluación y, concretamente, del papel que desempeñan los tests.

Como cualquier otro tipo de herramienta o instrumento, los tests pueden ser empleados al servicio de unas finalidades u otras, pero el compromiso debe ser la búsqueda constante e inequívoca del beneficio y no perjuicio de las personas, reaccionando ante cualquier tipo de impedimento que pueda limitar el ejercicio profesional honesto, libre e independiente. En definitiva, respetando el derecho de las personas a ser tratadas y evaluadas con justicia, equidad y responsabilidad, así como a conocer los resultados y las consecuencias que de la evaluación psicológica se puedan derivar. Como hemos podido ver, esta responsabilidad no se limita a la administración de tests con las garantías científicas necesarias, sino que implica el dominio de las teorías, los métodos y las técnicas que ha desarrollado la psicometría para comprender su funcionamiento, valorar su conveniencia con relación a los objetivos de la evaluación y, finalmente, evitar cualquier tipo de perjuicio que un uso inadecuado pudiera ocasionar. Por esa razón los psicólogos deben ser cuidadosos con su propia competencia, la clave de bóveda para un uso adecuado de los tests, con el objetivo de desarrollar los conocimientos y las habilidades necesarias que garanticen un profundo respeto por los derechos y la dignidad de las personas a quienes administran sus tests.

Capítulo II

Fiabilidad

Maite Barrios
Antoni Cosculluela

En el lenguaje cotidiano el término *fiabilidad* se asocia a algo que funciona de manera correcta. Nos fiamos de nuestro despertador si suena a la hora que se ha programado, de la báscula si nos proporciona sin error nuestro peso, incluso consideramos que contamos con un buen amigo si siempre nos apoya cuando lo necesitamos. Si el despertador, la báscula y nuestro amigo no se comportan de la manera "correcta", consideramos que no son fiables y en consecuencia decidimos que no podemos confiar en ellos.

En psicometría nos referimos a la fiabilidad como aquella propiedad que valora la consistencia y precisión de la medida. En consecuencia, si la medida toma valores consistentes y precisos, creemos que podemos confiar en los resultados obtenidos cuando se aplica un test. No obstante, sabemos que cualquier proceso de medida (se esté midiendo un objeto físico o un aspecto psicológico) se asocia a algún grado de error. La medida perfecta no existe. El estudio de la fiabilidad de un instrumento de medida debe permitir conocer hasta qué punto los resultados que se obtienen a partir de su aplicación están afectados por el error que se ha cometido al medir. Si el error es pequeño, podemos confiar en el resultado del test; si el error es grande, el proceso de medición deja de tener sentido. En este capítulo se trata el tema de la fiabilidad desde dos vertientes: por un lado, se presenta la fiabilidad desde la perspectiva de la teoría clásica de los tests (TCT), centrándonos en los test referidos a la norma (TRN), para, por otro lado, abordar el tema de la fiabilidad según los tests referidos al criterio (TRC).

Desde la perspectiva de la TCT se presenta lo que se entiende por error de medida y los diferentes tipos de error de medida que se pueden cometer al aplicar

un test. A continuación, se describe el modelo lineal propuesto por Spearman y cómo a partir de él se deriva el coeficiente de fiabilidad. Nos detendremos en cómo interpretar un coeficiente de fiabilidad, así como en las diferentes estrategias que se han ido desarrollando para calcularlo: test-retest, formas paralelas y consistencia interna. A continuación, tratamos tres de los factores que influyen en la fiabilidad (variabilidad de las puntuaciones obtenidas en el test, la longitud del test y las características de los ítems que lo componen). Para acabar con la TCT, se presentan dos procedimientos para valorar la puntuación verdadera de un sujeto: la estimación que asume la distribución normal del error aleatorio y la estimación a partir del modelo de regresión lineal.

Veremos una manera diferente de abordar la fiabilidad cuando los tests que se emplean son instrumentos cuyo objetivo es valorar la competencia de las personas en algún dominio concreto de conocimiento, los denominados TRC. Para contextualizar la fiabilidad en los TRC, en primer lugar, nos detenemos en los conceptos básicos que caracterizan este tipo de tests y, en segundo lugar, describimos los tres procedimientos más clásicos para abordar su fiabilidad: aquellos procedimientos que requieren dos aplicaciones del test para valorar la consistencia de la clasificación, aquellos que solo requieren una única aplicación y aquellos en los que entra en juego el papel de los evaluadores. En el último apartado del capítulo, se describen los métodos que más frecuentemente se utilizan para determinar el punto de corte que permite una mejor clasificación entre aquellos individuos que son competentes en el criterio de interés y aquellos que no lo son.

1. Concepto de fiabilidad según la teoría clásica

Según la teoría clásica de los tests, la fiabilidad de un test está relacionada con los errores de medida aleatorios presentes en las puntuaciones obtenidas a partir de su aplicación. Así, un test será más fiable cuantos menos errores de medida contengan las puntuaciones obtenidas por los sujetos a quienes se les aplica. Dicho de otro modo, la fiabilidad de un test será su capacidad para realizar medidas libres de errores.

1.1. El error de medida

Todo instrumento de medida debe garantizar, con más o menos rigor, que las medidas que obtenemos con su aplicación se corresponden con el verdadero nivel o valor de la característica evaluada. Así, si queremos medir la temperatura del agua del mar un día de un caluroso mes de agosto, necesitaremos un termómetro que nos permita obtener este dato. Si lo hacemos con el termómetro que compramos en unos grandes almacenes para medir la temperatura del agua de la bañera de casa, seguramente obtendremos un valor que será menos preciso que si lo hacemos con el termómetro que utiliza el servicio de meteorología para tomar estas medidas. En cualquier caso, seguramente tanto uno como otro termómetro medirán con un cierto grado de imprecisión, posiblemente más elevado en el primer caso que en el segundo, pero ninguno exento de una cierta desviación respecto a la verdadera temperatura del agua. Si la medida la hiciéramos utilizando un sofisticado instrumental cedido por la NASA, seguramente tendríamos bastantes más garantías de que la temperatura obtenida se corresponde con mucha más precisión con la verdadera.

Por lo tanto, cualquier proceso de medida de una característica de los objetos o de los sujetos lleva inherente un cierto error en su medición. Podemos encontrar instrumentos de medida con más o menos capacidad para minimizar estos errores, pero difícilmente podremos encontrar uno que los elimine del todo.

En nuestro ámbito de la psicología, donde las variables que medimos habitualmente son características propias de los sujetos, relacionadas con sus rasgos de personalidad, sus capacidades cognitivas, sus estados de ánimo, etc., y donde los instrumentos utilizados para la medición son generalmente los tests, aún resulta más evidente que las medidas que hacemos de estos atributos estarán también afectadas por ciertos errores. Esto provocará que las puntuaciones obtenidas con las administraciones de estos tests no se correspondan exactamente con los verdaderos niveles de los sujetos en la característica medida.

En cualquier caso, algunos de estos errores propios de toda medición pueden responder a factores sistemáticos que tendrán una posible causa en el propio proceso de medida, en el instrumento utilizado o en ciertas características de los objetos o sujetos medidos. Así, si el termómetro con el que medíamos la temperatura del agua del mar tiene un error de construcción que hace que siempre

mida un grado más del real, este error afectará por igual a toda medición realizada con él, y se podrá eliminar haciendo una buena calibración del aparato. Otros errores no tienen este componente sistemático, sino que son aleatorios, indeterminados y no responden a ningún factor que pueda ser conocido, y por lo tanto eliminado. Estos errores aleatorios son los que están implicados en el concepto de fiabilidad.

1.2. El coeficiente de fiabilidad y su interpretación

Desde la teoría clásica de los tests (TCT) de Spearman, se define el coeficiente de fiabilidad de un test $\rho_{xx'}$ como la correlación entre las puntuaciones obtenidas por un grupo de sujetos en dos formas paralelas del test.

Según la definición de formas paralelas de un test de la TCT, si un test tuviera una fiabilidad perfecta, las puntuaciones obtenidas por un sujeto en las dos formas paralelas del test deberían ser idénticas, y por lo tanto la correlación entre las puntuaciones de un grupo de sujetos en estas dos formas paralelas del test sería 1 ($\rho_{xx'} = 1$). Cualquier valor inferior a 1 se deberá a los errores aleatorios propios del instrumento de medida.

A partir de la definición anterior del coeficiente de fiabilidad, y teniendo en cuenta los supuestos de la TCT, también podemos expresar el coeficiente de fiabilidad como el cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas. En este sentido, el coeficiente de fiabilidad se puede interpretar como la proporción de varianza de las puntuaciones verdaderas $\left(\sigma_v^2\right)$ que hay en la varianza de las puntuaciones empíricas $\left(\sigma_x^2\right)$:

$$\rho_{xx'} = \frac{\sigma_v^2}{\sigma_x^2}$$

De la expresión anterior y de las consecuencias derivadas de la TCT podemos deducir fácilmente que este coeficiente de fiabilidad será igual a 1 menos la proporción de la varianza de los errores $\left(\sigma_e^2\right)$ que hay en la varianza de las puntuaciones empíricas.

$$\rho_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

Un índice directamente relacionado con el coeficiente de fiabilidad es el índice de fiabilidad (ρ_{xv}) que se define como la correlación entre las puntuaciones empíricas de un test y las puntuaciones verdaderas. Este índice de fiabilidad es igual a la raíz cuadrada del coeficiente de fiabilidad:

$$\rho_{xv} = \sqrt{\rho_{xx'}} = \frac{\sigma_v}{\sigma_x}$$

A la hora de interpretar el valor del coeficiente de fiabilidad no existe un criterio único y universalmente aceptado como adecuado. Evidentemente, valores cercanos a 0 denotarán una alta proporción de la varianza de los errores en la varianza de las puntuaciones empíricas, y por lo tanto, pondrán de manifiesto que el instrumento utilizado no es fiable, mientras que valores cercanos a 1 mostrarán una baja proporción de la varianza de los errores en la varianza de las puntuaciones empíricas y, en consecuencia, nos permitirán interpretar que el test utilizado es fiable. Ahora bien, el significado de esta varianza de error difiere con relación al tipo de estrategia que se ha utilizado para valorar la fiabilidad (estas estrategias se describen en los próximos apartados). Cohen y Swerlik (2009) proponen que si se ha utilizado la estrategia de test-retest, la varianza de error será debida fundamentalmente a las diferentes administraciones del test; si se ha utilizado la estrategia de formas paralelas, el error se puede atribuir a la construcción del test o a las diferentes administraciones, y si se ha valorado la fiabilidad a partir de la consistencia del test, la varianza de error puede deberse a la construcción del test.

Aparte de los casos extremos, la determinación del valor mínimo aceptable del coeficiente de fiabilidad depende de factores que pueden influir en este valor, como la longitud del test o el procedimiento empírico o la estrategia utilizada para su cálculo, tal como se ha comentado en el párrafo anterior. En cualquier caso, se han intentado establecer ciertos criterios generales que nos pueden servir de referencia. Así, en su texto clásico, Nunnally (1978) considera que el valor mínimo aceptable del coeficiente de fiabilidad estaría en 0,70, sobre todo en un contexto de investigación básica. En cambio, en un contexto aplicado, como el escolar o el clínico, es necesario que la fiabilidad sea más elevada, situándola por encima de 0,80 o 0,90. En estos ámbitos es necesario tener en cuenta que las consecuencias de la precisión de los instrumentos de medida utilizados pueden ser más decisivas para los sujetos evaluados (pensemos en los

tests de diagnóstico clínico, o en los de inteligencia en población infantil, para determinar la necesidad de clases especiales por los niños). Murphy y Davidshofer (2005) afirman que en cualquier contexto de evaluación una fiabilidad por debajo de 0,6 se consideraría baja e inaceptable. Kapplan y Saccuzo (2009) van algo más allá y sugieren que coeficientes de fiabilidad que oscilan entre 0,7 y 0,8 son suficientemente buenos para la mayoría de las ocasiones en las que los tests se utilizan para fines de investigación.

Otros autores consideran que un coeficiente de fiabilidad muy cercano a 1 puede significar que los ítems que componen el test son redundantes al evaluar ciertos elementos o factores del constructo medido, y por lo tanto no aportan información relevante respecto a otros elementos o factores de este constructo, lo que tampoco se puede considerar como adecuado.

Sin querer establecer criterios estrictos y teniendo en consideración todo lo que se ha expuesto hasta aquí, podríamos concluir que, en general, es posible interpretar como una fiabilidad adecuada valores del coeficiente de fiabilidad dentro del intervalo de 0,70 a 0,95.

1.3. Tipos de errores de medida

Hasta este momento solo nos hemos referido a un tipo de error: el error de medida, pero hay que mencionar que este no es el único error descrito en el ámbito de la psicometría, sino que también podemos hacer referencia al error de estimación, al error de sustitución y al error de predicción.

Estos errores están relacionados con las puntuaciones de los sujetos individualmente consideradas. Así, el error de medida es, tal como lo definiremos a continuación, la diferencia entre la puntuación obtenida por un sujeto en el test y su puntuación verdadera en la característica medida por este test. Ahora bien, si consideramos los errores no individualmente sino en relación con un grupo o muestra de sujetos, podemos obtener los denominados errores típicos, que son las desviaciones típicas de estos errores calculadas a partir de las puntuaciones de todos los sujetos de la muestra.

Por lo tanto, podemos definir más formalmente estos diferentes tipos de errores, sus errores típicos asociados y las fórmulas que los expresan.

– **Error de medida**. Definimos el error de medida como la diferencia entre la puntuación empírica de un sujeto (X) y su puntuación verdadera (V).

$$e = X - V$$

El *error típico de medida* es la desviación típica de los errores de medida, y lo podemos expresar como:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}}$$

– **Error de estimación de la puntuación verdadera**. El error de estimación de la puntuación verdadera se define como la diferencia entre la puntuación verdadera de un sujeto y su puntuación verdadera pronosticada mediante el modelo de la regresión (V').

$$e = V - V'$$

La desviación típica de estos errores de estimación se denomina *error típico de estimación* y se puede obtener con la siguiente expresión:

$$\sigma_{vx} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{\rho_{xx'}} = \sigma_e \sqrt{\rho_{xx'}}$$

– **Error de sustitución**. Se define el error de sustitución como la diferencia entre las puntuaciones de un sujeto en dos formas paralelas de un test o, dicho de otra manera, el error que se comete al sustituir la puntuación de un sujeto en un test (X_1), por la puntuación obtenida en una forma paralela de este mismo test (X_2).

$$e = X_1 - X_2$$

Se denomina *error típico de sustitución* a la desviación típica de los errores de sustitución, y lo podemos expresar del siguiente modo:

$$\sigma_{e(s)} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{2}$$

– **Error de predicción**. El error de predicción podemos definirlo como la diferencia entre la puntuación de un sujeto en un test (X_1) y la puntuación pronosticada en este test (X'_1) a partir de una forma paralela X_2. Sería el error que cometeríamos si sustituyéramos la puntuación de un sujeto en un test por la puntuación pronosticada a partir de una forma paralela de este test.

$$e = X_1 - X'_1$$

En este sentido, X'_1 será la puntuación pronosticada mediante la recta de regresión de X_1 sobre X_2, y la podemos expresar a partir del modelo lineal general adaptado a este contexto como:

$$X'_1 = \rho_{12} \frac{\sigma_1}{\sigma_2}\left(X_2 - \bar{X}_2\right) + \bar{X}_1$$

Definimos el *error típico de predicción* como la desviación típica de los errores de predicción, y lo podemos expresar como:

$$\sigma_{e(p)} = \sigma_x \sqrt{1 - \rho_{xx'}} \sqrt{1 + \rho_{xx'}}$$

2. Equivalencia de las medidas: Método de las formas paralelas

De la definición y las fórmulas del coeficiente de fiabilidad no se puede extraer directamente ningún procedimiento que nos permita calcular su valor para una determinada muestra de sujetos. Así, la estimación empírica del valor del coeficiente de fiabilidad hay que obtenerla mediante alguna estrategia que nos permita o bien comparar las puntuaciones de los mismos sujetos en dos administraciones del mismo test o en dos formas paralelas del test, o bien analizar las puntuaciones de un grupo de sujetos en los diferentes ítems del test.

De los procedimientos empíricos para la obtención del coeficiente de fiabilidad, el que se deriva directamente de la TCT es el llamado *método de las formas paralelas,* que consiste en el cálculo del coeficiente de correlación de Pearson entre las puntuaciones de una amplia muestra de sujetos representativa de la población diana del test, en dos formas paralelas de un test previamente obtenidas. Si, tal como se puede derivar de la definición de formas paralelas de un test, estas miden exactamente el mismo constructo, exactamente con la misma precisión, las diferencias que podremos observar entre las puntuaciones de unos mismos sujetos en las dos formas deben ser consecuencia de los errores de medida del test, y por lo tanto este procedimiento nos proporcionará un indicador adecuado de la magnitud de estos errores de medida, o sea, de la precisión o fiabilidad del test.

De hecho, este indicador también será representativo del grado de equivalencia de las dos formas paralelas del test, y en este sentido también recibe el nombre de *coeficiente de equivalencia*.

Su fórmula será la del coeficiente de correlación de Pearson aplicado a este caso:

$$r_{xx'} = r_{x_1x_2} = \frac{n\sum x_1x_2 - \sum x_1 \sum x_2}{\sqrt{\left[n\sum x_1^2 - \left(\sum x_1\right)^2\right]\left[n\sum x_2^2 - \left(\sum x_2\right)^2\right]}}$$

$r_{xx'}$: Coeficiente de fiabilidad del test.

$r_{x_1x_2}$: Coeficiente de correlación de Pearson.

x_1 y x_2: Puntuaciones obtenidas por los sujetos en cada una de las dos formas paralelas del test.

Hemos designado el coeficiente de fiabilidad del test con la notación de $r_{xx'}$ y no la de $\rho_{xx'}$ porque estamos obteniendo este coeficiente de fiabilidad para una muestra concreta de sujetos, y por lo tanto nos situamos a un nivel empírico y no a un nivel teórico, como hacíamos en los aparatos anteriores al presentar las bases y definiciones de la fiabilidad y de sus errores. Esta notación, que podemos definir como muestral (aplicada a una muestra concreta de sujetos), es la que seguiremos utilizando en los siguientes apartados, dado que en todos ellos se presentan los procedimientos empíricos para la obtención del coeficiente de fiabilidad y otros indicadores relacionados con él.

La mayor dificultad para la aplicación del método de las formas paralelas recae en la elaboración de estas dos versiones de un test. A menudo resulta realmente muy difícil construir dos tests que estén formados por ítems que puedan ser emparejados en función de su total equivalencia, como requiere el concepto y la definición de formas paralelas.

Los otros procedimientos para la obtención del coeficiente de fiabilidad se derivan de las dos vertientes que podemos considerar como inherentes al concepto de fiabilidad: la estabilidad temporal y la consistencia interna.

3. Estabilidad de las medidas: Método test-retest

Un instrumento de medida fiable debe proporcionar valores estables en diferentes medidas de los mismos sujetos secuencialmente obtenidas. Así, si me-

dimos un rasgo de personalidad supuestamente estable de unos mismos sujetos en dos diferentes ocasiones con el mismo test, el coeficiente de correlación entre sus puntuaciones será un buen indicador de esta estabilidad del test, y por lo tanto, de su fiabilidad.

Este método de obtención del coeficiente de fiabilidad de un test se denomina *método test-retest* y consiste en la aplicación del test a una misma muestra de sujetos en dos ocasiones diferentes. Se calcula a partir del valor del coeficiente de correlación de Pearson entre las puntuaciones de los sujetos en estas dos ocasiones. La fórmula es exactamente la misma aplicada en el apartado anterior para el caso de las formas paralelas, únicamente con la diferencia de que en el test-retest, x_1 y x_2 son las puntuaciones obtenidas por los sujetos en las dos administraciones del test.

La ventaja de este método es que no requiere dos formas diferentes del test (con todas las dificultades que esto implica) y su principal inconveniente se deriva en que hay que administrar dos veces el mismo test a los mismos sujetos. Este hecho puede suponer factores que distorsionen las puntuaciones de los sujetos en la segunda administración. Así, si los sujetos todavía recuerdan el contenido del test, seguramente su rendimiento se verá mejorado respecto a la primera administración. En este sentido, un factor crucial para una correcta aplicación de este método es determinar el intervalo temporal que hay que dejar entre las dos administraciones del test. Este intervalo temporal no puede ser ni demasiado corto como para provocar los efectos comentados anteriormente, ni demasiado amplio como para que se puedan dar cambios naturales (madurativos, evolutivos o circunstanciales) del rasgo o constructo medido que modifiquen las puntuaciones de los sujetos en esta segunda administración del test.

4. Consistencia interna

Tal como se ha dicho anteriormente, un instrumento de medida fiable se caracteriza por una elevada estabilidad temporal y por una adecuada consistencia interna. La consistencia interna hace referencia al grado en que cada una de las partes de las que se compone el instrumento es equivalente al resto. Este principio aplicado al caso de los tests vendrá determinado por el grado en el que

cada ítem, como parte básica constitutiva de este, muestra una equivalencia adecuada con el resto de los ítems, o sea, que mide con el mismo grado el constructo medido. Así, si hay una elevada equivalencia entre los ítems del test, es de suponer que las respuestas de los sujetos a estos diferentes ítems estarán altamente correlacionadas, y las diferentes partes en las que podamos dividir el test también mostrarán esta elevada covariación.

Por poner un ejemplo, pese a las evidentes diferencias que se pueden establecer con el caso de los tests, la consistencia interna de una cinta métrica queda garantizada si cada una de sus partes (supongamos los diferentes centímetros que la componen) es equivalente al resto. Así, podremos dividir la cinta en diferentes partes iguales (dos, tres etc.), y cada una de ellas medirá exactamente la misma distancia. Evidentemente, esta exactitud en la consistencia interna de la cinta métrica es prácticamente imposible de lograr en el caso de los tests, pero el ejemplo puede servir para situar adecuadamente el concepto de consistencia interna referido al caso de la construcción de instrumentos de medida en psicología.

4.1. Método de las dos mitades

Puede derivarse fácilmente, a partir de lo que se ha expuesto en el apartado anterior, que si dividimos un test en dos mitades, estas deben ser equivalentes para garantizar una adecuada consistencia interna. El grado de equivalencia de las dos mitades se puede evaluar calculando la correlación entre las puntuaciones de los sujetos en estas dos mitades. Así, la correlación entre las puntuaciones de un grupo de sujetos en las dos mitades en las que podemos dividir un test será un indicador del grado de consistencia interna de este, y por lo tanto de su fiabilidad. Este es el principio en el que se basa el *método de las dos mitades*, que presenta la ventaja respecto a los métodos anteriores de que solo requiere una sola aplicación del test a una muestra de sujetos.

A la hora de decidir cómo realizar esta partición del test en dos mitades, hay que tener en cuenta que si lo hacemos, por ejemplo, dejando los primeros ítems en una mitad y los últimos en la otra, pueden ponerse en juego factores que alteren la equivalencia del rendimiento de los sujetos en las dos mitades. Así, es conocido que los sujetos suelen prestar más atención a los primeros ítems de un test, con la consecuente mejora de su rendimiento, o a una mayor sinceridad en sus respuestas. Estos posibles factores incidirían en una falta de consistencia in-

terna del test, no producto de sus errores de medida aleatorios propios de la fiabilidad del test, sino de errores de medida sistemáticos independientes de esta fiabilidad, dado que una de las dos mitades del test se vería favorecida por un mejor o más cuidadoso rendimiento de los sujetos.

Para evitar factores como los comentados anteriormente, habitualmente se divide el test en dos mitades, dejando los ítems pares en una mitad y los impares en la otra. Con este procedimiento se evitan buena parte de estos factores y se garantiza de manera más probable la equivalencia entre las dos mitades.

4.1.1. Spearman-Brown

Como se comentará con más detalle en los próximos apartados, el número de ítems que componen un test incide en su fiabilidad. Así, siendo constantes otros factores, cuantos más ítems contiene un test más elevada es su fiabilidad. Este efecto de la longitud de un test sobre el coeficiente de fiabilidad hay que tenerlo presente al aplicar el método de las dos mitades. Por lo tanto, si calculamos el coeficiente de correlación entre el total de las puntuaciones de los sujetos en los ítems pares por un lado y por otro el total de sus puntuaciones en los ítems impares, y a partir de este coeficiente de correlación cuantificáramos la fiabilidad del test, este valor estaría negativamente sesgado, dado que lo calcularíamos a partir de la correlación entre la mitad del número total de ítems del test. Este hecho supone que hay que realizar una corrección de este coeficiente de correlación para obtener el coeficiente de fiabilidad de la totalidad del test. Esta corrección se denomina de Spearman-Brown y es un caso concreto de la fórmula del mismo nombre que se aplica para obtener la fiabilidad de un test una vez este se ha alargado o acortado, añadiendo o eliminando una determinada cantidad de ítems:

La fórmula para la obtención del coeficiente de fiabilidad de un test a partir del método de las dos mitades con la corrección de Spearman-Brown es:

$$r_{xx'} = \frac{2r_{pi}}{1 + r_{pi}}$$

$r_{xx'}$: Coeficiente de fiabilidad del test.

r_{pi}: Coeficiente de correlación de Pearson entre el sumatorio de las puntuaciones de los ítems pares y las de los ítems impares.

Ejemplo

En la tabla siguiente tenemos las puntuaciones de ocho sujetos en un test de seis ítems dicotómicos:

Tabla 1

Sujetos	Ítems					
	1	2	3	4	5	6
A	1	1	1	1	0	1
B	0	1	1	1	1	0
C	1	1	0	1	1	0
D	1	1	1	1	1	1
E	1	1	1	1	1	1
F	0	1	1	0	0	0
G	0	1	1	0	1	0
H	1	0	1	0	0	0

Para obtener el coeficiente de fiabilidad del test por el método de las dos mitades, en primer lugar calculamos el sumatorio de las puntuaciones en los ítems pares para cada sujeto por un lado, y por otro el sumatorio de sus ítems impares, y obtenemos el coeficiente de correlación entre estas dos distribuciones de valores:

Tabla 2

Sujetos	Ítems pares	Ítems impares
A	3	2
B	2	2
C	2	2
D	3	3
E	3	3

Sujetos	Ítems pares	Ítems impares
F	1	1
G	1	2
H	0	2

Aplicando la fórmula del coeficiente de correlación de Pearson entre los ítems pares y los impares obtenemos un coeficiente igual a 0,62 $\left(r_{pi} = 0,62\right)$.

Una vez calculado este valor, aplicamos la fórmula de Spearman Brown para obtener el coeficiente de fiabilidad del test:

$$r_{xx'} = \frac{2r_{pi}}{1+r_{pi}} = \frac{2 \times 0,62}{1+0,62} = 0,76$$

El coeficiente de fiabilidad del test es de 0,76.

4.1.2. Rulon

La fórmula de Rulon (1939) para calcular la fiabilidad de un test también parte de su división en dos mitades. Se basa en el supuesto de que si las dos mitades son paralelas, las puntuaciones de los sujetos en cada una de ellas solo pueden diferir como consecuencia de los errores aleatorios. Por lo tanto, la varianza de las diferencias entre estas dos mitades será una estimación de la varianza de los errores y podremos sustituir la varianza de los errores de la fórmula del coeficiente de fiabilidad derivada de la TCT por la varianza de las diferencias.

Así, la fórmula de Rulon es la siguiente:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2}$$

S_d^2 : Varianza de las diferencias entre las puntuaciones de los sujetos en las dos mitades del test.

S_x^2 : Varianza de las puntuaciones totales de los sujetos en el test.

Ejemplo

Si aplicamos la fórmula de Rulon al ejemplo del apartado anterior, necesitaremos calcular:

Tabla 3

Sujetos	Ítems pares	Ítems impares	Diferencia P-I	Total
A	3	2	1	5
B	2	2	0	4
C	2	2	0	4
D	3	3	0	6
E	3	3	0	6
F	1	1	0	2
G	1	2	-1	3
H	0	2	-2	2

Después de calcular la varianza de las diferencias $\left(S_d^2 = 0,6875\right)$ y la varianza de las puntuaciones totales $\left(S_x^2 = 2,25\right)$, podemos aplicar la fórmula de Rulon:

$$r_{xx'} = 1 - \frac{S_d^2}{S_x^2} = 1 - \frac{0,6875}{2,25} = 0,69$$

El coeficiente de fiabilidad del test obtenido con la fórmula de Rulon es de 0,69.

4.1.3. Gutman-Flanagan

Tanto Flanagan (1937) como Gutman (1945) obtuvieron una fórmula equivalente a la de Rulon a partir de las varianzas de los ítems pares e impa-

res. Se basa en el mismo principio que el anterior, pero resulta más sencilla de obtener:

$$r_{xx'} = 2\left(1 - \frac{(S_p^2 + S_i^2}{S_x^2}\right)$$

S_p^2 : Varianza de las puntuaciones de los sujetos en los ítems pares del test.

S_i^2 : Varianza de las puntuaciones de los sujetos en los ítems impares del test.

S_x^2 : Varianza de las puntuaciones totales de los sujetos en el test.

Ejemplo

Aplicando la fórmula de Gutman-Flanagan al ejemplo anterior, tenemos:

Tabla 4

Sujetos	Ítems pares	Ítems impares	Total
A	3	2	5
B	2	2	4
C	2	2	4
D	3	3	6
E	3	3	6
F	1	1	2
G	1	2	3
H	0	2	2

Calculando las diferentes varianzas obtenemos:

Varianza de las puntuaciones en los ítems pares: $S_p^2 = 1,11$

Varianza de las puntuaciones en los ítems impares: $S_i^2 = 0,36$

Varianza de las puntuaciones totales en el test: $S_x^2 = 2,25$

$$r_{xx'} = 2\left(1 - \frac{S_p^2 + S_i^2}{S_x^2}\right) = 2\left(1 - \frac{1,11 + 0,36}{2,25}\right) = 0,69$$

Como podemos observar, el valor del coeficiente de fiabilidad, calculado a partir de la expresión de Gutman-Flanagan, es exactamente igual al obtenido con la fórmula de Rulon, como no podía ser de otra manera, dado que, como hemos dicho, las dos fórmulas son equivalentes. Tanto una como la otra proporcionan un coeficiente de fiabilidad del test igual a 0,69.

4.2. Covariación entre los ítems

Si, tal como se ha comentado anteriormente, la consistencia interna de un test hace referencia al grado en el que cada una de las partes o ítems de los que se compone es equivalente al resto, la covariación entre estos ítems también nos proporcionará un adecuado indicador de esta consistencia interna. De hecho, este procedimiento es una extensión de los procedimientos anteriores de la división del test en dos mitades, al caso límite de dividirlo en tantas partes como ítems lo componen. Así, cada ítem representará una parte equivalente del conjunto de todos ellos, es decir, del test o escala total. Del mismo modo que las dos partes del test deben mantener una elevada correlación entre ellos para garantizar la misma consistencia interna del conjunto, cada ítem también ha de mostrar una covariación adecuada con el resto de los ítems.

4.2.1. Coeficiente alfa de Cronbach

El coeficiente alfa (α) de Cronbach (1951) expresa la consistencia interna de un test a partir de la covariación entre sus ítems. Cuanto más elevada sea la proporción de la covariación entre estos ítems respecto a la varianza total del test, más elevado será el valor del coeficiente alfa (α) de Cronbach, y más elevada su fiabilidad.

Existen diferentes fórmulas para obtener el valor del coeficiente α, la más ampliamente utilizada de las cuales es la que se deriva del cálculo de las varianzas de cada ítem y de la varianza de las puntuaciones totales en el test.

Esta fórmula es la siguiente:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{j=1}^{n} S_j^2}{S_x^2}\right]$$

n: Número de ítems del test.

$\sum_{j=1}^{n} S_j^2$: Sumatorio de las varianzas de los *n* ítems.

$S_x^2 =$: Varianza de las puntuaciones totales en el test.

En nuestro ejemplo:

Tabla 5

Sujetos	**Ítems**						
	1	2	3	4	5	6	x
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2
Varianzas	0,234	0,109	0,109	0,234	0,234	0,234	2,25

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{j=1}^{n} S_j^2}{S_x^2}\right] = \frac{6}{5}\left(1 - \frac{0,234 + 0,109 + 0,109 + 0,234 + 0,234 + 0,234}{2,25}\right) = 0,583$$

Por otro lado, la fórmula del coeficiente alfa que se deriva directamente de la covarianza entre los diferentes ítems viene expresada por:

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

n: Número de ítems del test.

$\sum\sum \mathrm{cov}(j,k) =$: Sumatorio de las covarianzas de los *n* ítems.

S_x^2: Varianza de las puntuaciones en el test.

Así, en el ejemplo expuesto anteriormente, calcularemos la varianza de las puntuaciones totales en el test (*x*), y las covarianzas entre los diferentes ítems:

Tabla 6

Sujetos	**Ítems**						
	1	2	3	4	5	6	x
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2

Las covarianzas entre los 6 ítems son:

Tabla 7

	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Ítem 6
Ítem 1		-0,047	-0,047	0,109	-0,016	0,141
Ítem 2	-0,047		-0,016	0,078	0,078	0,047
Ítem 3	-0,047	-0,016		-0,047	-0,047	0,047
Ítem 4	0,109	0,078	-0,047		0,109	0,141
Ítem 5	-0,016	0,078	-0,047	0,109		0,016
Ítem 6	0,141	0,047	0,047	0,141	0,016	

y su sumatorio:

$$\sum\sum \text{cov}(j,k) = -0,047 + -0,047 + 0,109 + \ldots\ldots + 0,047 + 0,141 + 0,016 = 1,094$$

mientras que la varianza de las puntuaciones totales: $S_x^2 = 2,25$

Por lo tanto:

$$\alpha = \frac{n}{n-1}\left[\frac{\sum\sum_{j\neq k}^{n}\text{cov}(j,k)}{S_x^2}\right] = \frac{6}{5}\left(\frac{1,094}{2,25}\right) = 0,583$$

La fórmula del coeficiente α también se puede expresar en función del cociente entre la media de las covarianzas y la media de las varianzas de los diferentes ítems del test. Este cociente, que designamos como r_1, constituye una estimación de la fiabilidad de cada ítem. En este sentido, la fórmula del coeficiente α a partir de r_1 es una aplicación de la corrección de Spearman-Brown, que hemos comentado para el caso de las dos mitades, a partir de la estimación de la fiabilidad de cada ítem, teniendo en cuenta que si tenemos *n* ítems es como si hubiéramos alargado *n* veces el ítem inicial.

$$\alpha = \frac{n(r_1)}{1+(n-1)r_1}$$

Para nuestro ejemplo, la media de las covarianzas es 1,094/30, mientras que la media de las varianzas es 1,154/6.

Por lo tanto, $r_1 = \frac{1{,}094/30}{1{,}154/6} = 0{,}189$ y el valor de α:

$$\alpha = \frac{n(r_1)}{1+(n-1)r_1} = \frac{6\times 0{,}189}{1+(5\times 0{,}189)} = 0{,}583$$

Como podemos observar, y como no podía ser de otra manera, todas las diferentes fórmulas de cálculo del coeficiente α de Cronbach aplicadas a los datos de nuestro ejemplo nos proporcionan el mismo valor.

4.2.2. Inferencias sobre α

Una vez hemos obtenido el valor del coeficiente alfa de Cronbach para una muestra determinada de sujetos, podemos estar interesados en comprobar su significación estadística, o en determinar entre qué valores puede fluctuar este coeficiente en la población. Por otro lado, también puede interesarnos comparar dos coeficientes alfas obtenidos en dos muestras independientes, o en la propia muestra, para determinar si la diferencia entre ellos es estadísticamente significativa.

Contraste para un solo coeficiente

Kristof (1963) y Feldt (1965) propusieron un estadístico de contraste para comprobar si un determinado valor del coeficiente alfa puede ser compatible con un cierto valor poblacional. Así, podemos analizar si este valor de alfa es estadísticamente significativo, esto es, si podemos descartar la hipótesis de que su valor en la población es cero, o si este valor difiere significativamente o no de un determinado valor previamente fijado en la población.

El estadístico de contraste es:

$$F = \frac{1-\alpha}{1-\hat{\alpha}}$$

N: Número de sujetos.

n: Número de ítems.

α: Valor de alfa en la población.

$\hat{\alpha}$: Valor de alfa calculado en la muestra.

Y se distribuye según una distribución *F* de Snedecor con $(N - 1)$ y $(n - 1)(N\text{-}1)$ grados de libertad.

Podemos aplicar este estadístico de contraste a nuestro ejemplo del test de seis ítems que nos ha dado un valor de alfa de 0,583, obtenido en una muestra de ocho sujetos.

La hipótesis nula que plantearemos es que este coeficiente alfa es igual a cero en la población, caso más habitual y que supone la no significación estadística de este coeficiente, mientras que la alternativa supondrá la desigualdad respecto al valor cero, y por lo tanto su significación estadística.

Los pasos que se deberán seguir en este contraste serán:

Hipótesis nula: $\alpha = 0$

Hipótesis alternativa: $\alpha \neq 0$

Cálculo del estadístico de contraste:

$$F = \frac{1-\alpha}{1-\hat{\alpha}} = \frac{1-0}{1-0,583} = 2,398$$

Los valores críticos de la distribución *F* de Snedecor con 7 (*N*-1) y 35 ((*n*-1)(*N*-1)) grados de libertad, para un nivel de confianza del 95% y contraste bilateral, son:

$$F_{0,975(\ 7,35)} \approx 2,62 \text{ y } F_{0,025(\ 7,35)} \approx 0,23$$

$$F_{0,025(7,35)} = \frac{1}{F_{0,975(35,7)}} \approx \frac{1}{4,31} = 0,23$$

Como el valor del estadístico de contraste obtenido (2,398) se encuentra dentro del intervalo comprendido entre los valores críticos 2,62 y 0,23, aceptamos la hipótesis nula y podemos concluir que, a partir de nuestros datos y con un nivel de confianza del 95%, no tenemos evidencia suficiente para descartar que el valor del coeficiente alfa en la población es cero, por lo que este coeficiente no es estadísticamente significativo.

Como derivación sencilla de lo que se ha expuesto en este apartado, podemos también determinar el intervalo de confianza para el valor del coeficiente alfa obtenido. En este sentido, solo hay que sustituir, en la fórmula del estadístico de contraste, los valores críticos de la distribución *F* y aislar los valores de α:

$$\frac{1-\alpha}{1-0,583} \leq 2,62$$

$$\alpha \geq 1 - 2,62(1-0,583) = -0,09$$

$$\frac{1-\alpha}{1-0,583} \geq 0,23$$

$$\alpha \leq 1 - 0,23(1-0,583) = 0,90$$

La interpretación de estos valores irá en el sentido de considerar que, con un nivel de confianza del 95%, los valores del coeficiente alfa en la población estarán comprendidos entre –0,09 y 0,90. Una vez establecido este intervalo confidencial, podríamos resolver la aceptación o no de cualquier valor del coeficiente en la hipótesis nula del contraste correspondiente. Así, si el valor del coeficiente alfa poblacional planteado en la hipótesis nula cae dentro del intervalo de confianza, no podemos descartar la certeza de esta hipótesis nula, mientras que si no cae, podremos descartarla y aceptar la hipótesis alternativa para el nivel de confianza establecido.

En nuestro ejemplo, como el valor de cero del coeficiente está comprendido en el intervalo (–0,09 - 0,90) no podemos rechazar la hipótesis nula, tal como ya hemos visto anteriormente. Se deriva directamente de lo anterior el hecho de que si un coeficiente alfa empíricamente obtenido no se encuentra comprendido en el intervalo de confianza construido, queda determinada su significación estadística sin necesidad de realizar el contraste para una hipótesis nula igual a cero.

Contraste para dos coeficientes en muestras independientes

También podemos estar interesados en comprobar si dos coeficientes alfa obtenidos en muestras diferentes de sujetos son iguales o no. Para responder a esta cuestión aplicaremos un contraste para dos coeficientes en muestras independientes. Feldt (1969) propuso el estadístico *w*, que permite determinar si la diferencia entre los dos coeficientes es estadísticamente significativa:

$$w = \frac{1 - \hat{\alpha}_1}{1 - \hat{\alpha}_2}$$

$\hat{\alpha}_1$ y $\hat{\alpha}_2$: Coeficientes obtenidos en cada una de las dos muestras.

N_1 y N_2: Dimensiones de estas muestras.

donde *w* se distribuye según una F de Snedecor con (N_1 – 1) y (N_2 – 1) grados de libertad.

Podemos realizar este contraste para nuestro ejemplo con un nivel de confianza del 95%, suponiendo que se ha administrado el mismo test a una muestra de 10 sujetos y que hemos obtenido un coeficiente alfa de 0,65. Realizaremos el contraste siguiendo los siguiendo pasos:

Hipótesis nula: $\hat{\alpha}_1 = \hat{\alpha}_2$

Hipótesis alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

Cálculo del estadístico de contraste: $w = \frac{1 - \hat{\alpha}_1}{1 - \hat{\alpha}_2} = \frac{1 - 0{,}583}{1 - 0{,}65} = 1{,}19$

Los valores críticos de la distribución *F* de Snedecor con 7 (N_1) y 9 (N_2) grados de libertad, para un nivel de confianza del 95% y contraste bilateral, son:

$$F_{0{,}975(7,9)} = 4{,}20 \text{ y } F_{0{,}025(7,9)} = 0{,}21$$

Como el valor del estadístico de contraste obtenido (1,19) cae dentro del intervalo comprendido entre los valores críticos (0,21 - 4,20), no tenemos suficientes evidencias para rechazar la hipótesis nula, y por lo tanto debemos

concluir que la diferencia entre los dos coeficientes no es estadísticamente significativa[1].

Contraste para dos coeficientes en muestras dependientes

Es habitual que los dos coeficientes alfa obtenidos se hayan calculado a partir de la misma muestra de sujetos. Menos frecuente es que se hayan obtenido a partir de dos muestras de sujetos relacionados entre ellos por algún criterio de emparejamiento (por ejemplo, parejas de gemelos, padre-madre). No obstante, tanto en uno como en el otro supuesto denominamos el diseño como muestras dependientes. Sería el caso, por ejemplo, de aplicar un diseño experimental de medidas repetidas y administrar el mismo test a un solo grupo de sujetos en dos ocasiones diferentes. En este sentido, podríamos comparar los dos coeficientes alfa obtenidos para determinar la posible diferencia estadísticamente significativa entre ellos.

Feldt (1980) propuso un estadístico de contraste para comparar dos coeficientes α obtenidos en muestras dependientes:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12})}}$$

t: Se distribuye según una distribución t de Student con $N - 2$ grados de libertad.

$\hat{\alpha}_1$ y $\hat{\alpha}_2$: : Valores de los dos coeficientes alfa.

N: Número de sujetos de la muestra.

r_{12}: Correlación entre las puntuaciones de los sujetos en las dos administraciones del test.

Para ilustrar la aplicación de este contraste podemos considerar que el test utilizado en los apartados anteriores lo hemos administrado de nuevo en una segunda ocasión a la misma muestra de ocho sujetos.

1. Woodruff y Feldt (1986) hicieron extensivo el contraste anterior al caso de más de dos coeficientes comparados simultáneamente. No nos extenderemos en esta cuestión, dado que su aplicación no es tan frecuente, pero remitimos al lector interesado a los textos de Muñiz (2003) y Barbero, Vila y Suárez (2003), en los que se puede encontrar su desarrollo claramente explicado.

En la tabla siguiente tenemos las puntuaciones de estos sujetos en las dos ocasiones en las que se les ha administrado el test.

Tabla 8

	Ocasión 1							**Ocasión 2**						
	Ítems							**Ítems**						
Sujetos	1	2	3	4	5	6	x_1	1	2	3	4	5	6	x_2
A	1	1	1	1	0	1	5	1	1	0	1	0	1	4
B	0	1	1	1	1	0	4	0	1	0	1	1	0	3
C	1	1	0	1	1	0	4	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6	1	1	0	1	1	1	5
E	1	1	1	1	1	1	6	1	1	0	1	1	1	5
F	0	1	1	0	0	0	2	0	1	0	0	0	0	1
G	0	1	1	0	1	0	3	0	1	0	0	1	0	2
H	1	0	1	0	0	0	2	1	0	0	0	0	0	1

A partir de estos datos, calculamos el coeficiente alfa para la segunda administración del test ($\hat{\alpha}_2 = 0{,}718$), y la correlación entre las puntuaciones totales de los sujetos en estas dos ocasiones (r_{12} = 98).

El contraste para determinar la posible diferencia estadísticamente significativa entre los dos coeficientes alfa, con un nivel de confianza del 95%, seguirá los pasos siguientes:

Hipótesis nula: = $\hat{\alpha}_1 = \hat{\alpha}_2$

Hipótesis alternativa: $\hat{\alpha}_1 \neq \hat{\alpha}_2$

Cálculo del estadístico de contraste:

$$t = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12})}} = \frac{(0{,}583 - 0{,}718)\sqrt{8-2}}{\sqrt{4(1-0{,}583)(1-0{,}718)(1-0{,}98)}} = -3{,}41$$

Los valores críticos de la distribución *t* de Student con 6 (*N* – 2) grados de libertad, para un nivel de confianza del 95% y contraste bilateral, son:

$$t_{0,975(6)} = 2,447 \text{ y } t_{0,025(6)} = -2,447$$

Como el estadístico de contraste obtenido (–3,41) queda fuera del intervalo entre los valores críticos (–2,447 - 2,447), podemos rechazar la hipótesis nula y aceptar la alternativa, y concluir que con un nivel de confianza del 95% la diferencia entre los dos coeficientes alfa es estadísticamente significativa.

Como en el caso de dos muestras independientes, Woodruff y Feldt (1986) también hicieron extensivo el contraste anterior, además de dos coeficientes comparados simultáneamente en muestras dependientes. Se puede consultar su desarrollo en los mismos textos citados para muestras independientes.

4.2.3. Kuder-Richardson

Algunos años antes de que Cronbach propusiera el coeficiente alfa como indicador de la consistencia interna de un test, Kuder y Richardson (1937) presentaron dos fórmulas de cálculo de este indicador, que de hecho son casos particulares de α cuando los ítems son dicotómicos. Estas dos fórmulas son conocidas como KR_{20} y KR_{21}.

Cuando los ítems de un test son dicotómicos y se codifican las dos alternativas de respuesta posibles como 0 y 1, la varianza de un ítem es igual a la proporción de ceros para la proporción de unos. Si el test es de rendimiento y las respuestas a los diferentes ítems son correctas o incorrectas, habitualmente se codifica con un 1 las respuestas correctas y con un 0 las incorrectas. En este caso, la varianza del ítem será igual a la proporción de sujetos que aciertan el ítem (p_j) por la proporción de sujetos que no lo aciertan (q_j). Igualmente, si el test es de personalidad y no hay respuestas correctas ni incorrectas, pero se codifica con un 1 los sujetos que responden "SÍ" y con un 0 los que responden "NO", la varianza del ítem será la proporción de sujetos que responden "SÍ" (p_j) por la proporción de sujetos que responden "NO" (q_j). En los dos casos $S_j^2 = p_j q_j$.

Teniendo en cuenta esta igualdad, la fórmula del KR_{20} simplemente sustituye en la del coeficiente α de Cronbach el sumatorio de las varianzas de los ítems por el sumatorio de los productos p_j por q_j:

$$\mathrm{KR}_{20} = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} p_j q_j}{S_x^2}\right)$$

En el supuesto de que todos los ítems tuvieran la misma dificultad, o de que el número de sujetos que responden "SÍ" se mantuviera constante para todos los ítems, el producto p_j por q_j sería igual para todos ellos, y su sumatorio sería igual a la media del test menos esta media al cuadrado dividida por el número de ítems (n). Esta nueva igualdad permite reformular el KR_{20} para el caso de que todos los ítems tengan la misma dificultad o el número de sujetos que responden "SÍ" se mantenga constante para todos los ítems. Esta reformulación es el KR_{21}.

$$\mathrm{KR}_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{X} - \bar{X}^2 / n}{S_x^2}\right)$$

Si aplicamos estas fórmulas a nuestro ejemplo, obtenemos los resultados siguientes:

Tabla 9

Sujetos	**Ítems**						
	1	2	3	4	5	6	*x*
A	1	1	1	1	0	1	5
B	0	1	1	1	1	0	4
C	1	1	0	1	1	0	4
D	1	1	1	1	1	1	6
E	1	1	1	1	1	1	6
F	0	1	1	0	0	0	2

Sujetos	Ítems						
	1	2	3	4	5	6	*x*
G	0	1	1	0	1	0	3
H	1	0	1	0	0	0	2
$p_i q_i$	0,234375	0,109375	0,109375	0,234375	0,234375	0,234375	

$$\sum p_j q_j = 0,2343375 + 0,109375 + 0,109375 + 0,234375$$
$$+0,234375 + 0,234375 = 1,15625$$

$$S_x^2 = 2,25$$

$$\bar{X} = 4$$

$$\mathrm{KR}_{20} = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} p_j q_j}{S_x^2}\right) = \frac{6}{5}\left(1 - \frac{1,15625}{2,25}\right) = 0,583$$

$$\mathrm{KR}_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{X} - \bar{X}^2/n}{S_x^2}\right) = \frac{6}{5}\left(1 - \frac{4 - 4^2/6}{2,25}\right) = 0,489$$

Como podemos comprobar, el KR_{20} proporciona un resultado idéntico al del coeficiente alfa, mientras que el KR_{21} da un resultado inferior, dado que en nuestro ejemplo no todos los ítems tienen la misma dificultad. En este sentido, el cálculo del KR_{21} no sería adecuado para este caso.

5. Factores que afectan a la fiabilidad

La fiabilidad de un test depende de factores como la variabilidad de las puntuaciones del test, el número total de ítems del test o las características de los ítems que lo componen. A continuación, se tratarán estos tres aspectos: variabilidad, longitud y características de los ítems.

a) Variabilidad

En los apartados previos se ha abordado la fiabilidad a partir del cálculo del coeficiente de correlación entre dos tests paralelos, entre dos administraciones del test en dos momentos temporales diferentes o entre diferentes partes del test. Sin embargo, hay que tener en cuenta que el coeficiente de correlación es sensible al rango y variabilidad de los datos. Lo que se observa es que cuando mantenemos el resto de los factores constantes, si se aumenta la variabilidad de los datos, el coeficiente de correlación aumenta. Por esta razón, en aquellos casos en los que exista una alta variabilidad en las puntuaciones del test, el coeficiente de fiabilidad será mayor. De esto se desprende que un test no tiene un coeficiente de fiabilidad único y fijo, sino que depende de las características de la muestra sobre la que se calcula. Así, por el contrario, si la muestra es homogénea y las puntuaciones empíricas que se obtienen presentan una baja variabilidad, el coeficiente de fiabilidad será menor.

En este punto vale la pena recordar las palabras de Crocker y Algina (1986) cuando dicen que no se puede afirmar que un test es fiable o no, sino que la fiabilidad es una propiedad de las puntuaciones obtenidas en el test a partir de una muestra particular de individuos.

b) Longitud

Otro de los factores que afectan a la fiabilidad es la longitud del test. Así, la fiabilidad depende del número de ítems que presente el test. La lógica de esta afirmación subyace en que cuantos más ítems se utilicen para medir un constructo, mejor podrá ser valorado este y menor será el error de medida que se cometerá al valorar la puntuación verdadera del sujeto. Por ello, siempre que se aumente el número de ítems de un test (siempre que estos sean ítems representativos del constructo), la fiabilidad aumentará. Para saber la fiabilidad de un test en caso de que aumente o disminuya su número de ítems, se utiliza la fórmula de Spearman Brown, también conocida como profecía de Spearman Brown.

$$R_{xx} = \frac{kr_{xx}}{1 + (k-1)r_{xx}}$$

R_{xx}: Nuevo coeficiente de fiabilidad del test alargado o acortado.

r_{xx}: Coeficiente de fiabilidad del test original.

k: Número de veces que se alarga o se acorta el test.

De este modo, k vendrá dado por el cociente entre el número de ítems finales (n_f) del test dividido por el número de ítems iniciales (n_i) del test:

$$k = \frac{n_f}{n_i}$$

Si se añaden ítems a un test, k siempre será superior a 1, mientras que si se acorta el test (eliminamos ítems a los ya existentes) k será inferior a 1.

Ejemplo

Si un test de 25 ítems presenta una fiabilidad de 0,65 y le añadimos 10 ítems paralelos, ¿cuál será su fiabilidad?

$$k = \frac{35}{25} = 1,4$$

k indica que hay que alargar 1,4 veces la longitud del test. Si se sustituye este valor en la fórmula:

$$R_{xx} = \frac{1,4 \cdot 0,65}{1 + (1,4 - 1) \cdot 0,65} = 0,72$$

Se obtiene que el nuevo coeficiente de fiabilidad del test alargado será de 0,72.

También la pregunta anterior se puede invertir y plantearse cuántos ítems debería tener el test para lograr una determinada fiabilidad. En este caso, habría que aislar k de la fórmula:

$$k = \frac{R_{xx}(1 - r_{xx})}{r_{xx}(1 - R_{xx})}$$

Efectivamente, si ahora la pregunta fuera cuántos ítems hay que añadir para conseguir una fiabilidad de 0,72 a un test de 25 ítems que presenta una fiabilidad de 0,65, se aplicaría la fórmula anterior para conocer cuántas veces habría que aumentar el test:

$$k = \frac{0,72 \cdot (1 - 0,65)}{0,65 \cdot (1 - 0,72)} = 1,4$$

Lo que permitirá saber el número de ítems que hay que añadir:

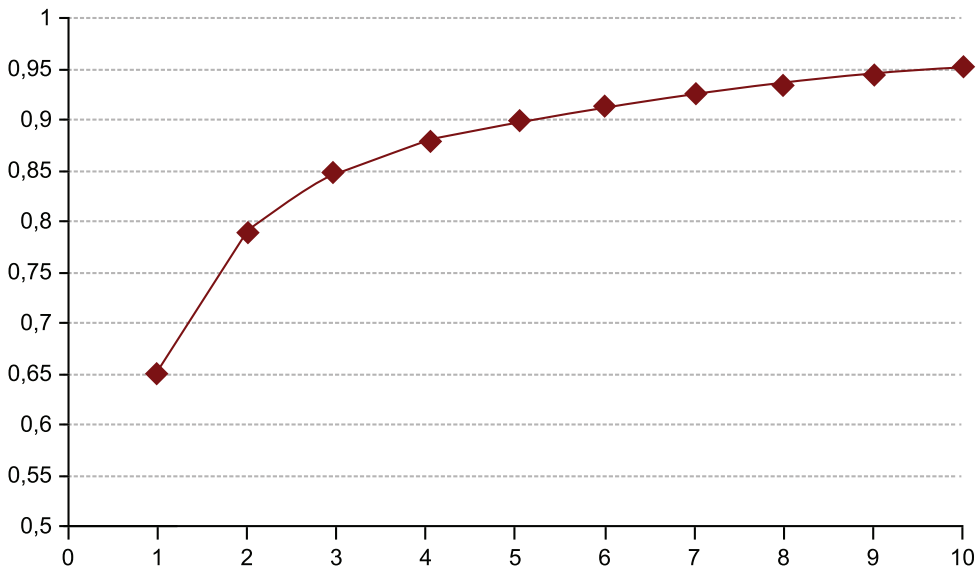
$$k \cdot n_i - n_i$$

$$1,4 \cdot 25 - 25 = 10$$

El test final tendría 35 ítems (1,4 × 25 = 35), por lo que habría que añadir 10 ítems a los 25 iniciales para conseguir una fiabilidad de 0,72.

Hay que tener presente dos aspectos: el primero es que a pesar de que la fiabilidad de un test aumentará siempre que aumentemos el número de ítems, este aumento no es directamente proporcional. Para ver el efecto del aumento que se produce en la fiabilidad en diferentes valores de *k*, hay que fijarse en la figura 1, en la que en las abscisas se representan los diferentes valores de *k* (el número de veces que se ha alargado el test) y en las ordenadas el coeficiente de fiabilidad que se obtendría.

Figura 1. Relación entre el coeficiente de fiabilidad y el aumento de ítems en un test



El segundo aspecto que hay que tener en cuenta es que, aunque siempre podemos conseguir un aumento de la fiabilidad aumentando el número de ítems del test, se deben valorar los aspectos de fatiga que supone responder a un ins-

trumento con muchos ítems. Por ello, se puede pensar en aumentar el número de ítems si el test original tiene relativamente pocos ítems y una fiabilidad que no llega a ser del todo adecuada. Pero si el test ya presenta un número considerable de ítems o su fiabilidad dista mucho ser adecuada, quizá habría que seleccionar otro test para medir el constructo o construir un nuevo test.

c) Características de los ítems

Cada ítem del test contribuye de manera específica a la fiabilidad o consistencia interna del test. Una manera de comprobarlo es calcular el coeficiente alfa de Cronbach eliminando del cálculo la puntuación del ítem. Si el ítem contribuye de manera positiva a la consistencia interna del test, al eliminarlo del test, el valor del coeficiente alfa de Cronbach se verá alterado a la baja (si eliminamos el ítem, el test pierde consistencia interna). Al contrario, si observamos que al eliminar el ítem el coeficiente alfa de Cronbach aumenta, esto indicará que el ítem no contribuye de manera positiva a la consistencia interna.

Por ejemplo, si un test con seis ítems presenta una consistencia interna de 0,76 y se calcula de nuevo la consistencia interna eliminando la puntuación del ítem en cada caso, habría que valorar que en todos los casos la consistencia interna fuera inferior a la del conjunto del test.

En la tabla siguiente se muestra un ejemplo:

Tabla 10

Ítems	Alfa de Cronbach sin el ítem
Ítem 1	0,72
Ítem 2	0,74
Ítem 3	0,80
Ítem 4	0,71
Ítem 5	0,70
Ítem 6	0,75

En la tabla anterior, se observa que cuando se elimina un ítem el valor del coeficiente alfa de Cronbach disminuye (ítems 1, 2, 4, 5 y 6), lo que indica que

el ítem contribuye de manera favorable a la consistencia interna del test. No obstante, cuando se elimina el ítem 3 el valor del coeficiente alfa de Cronbach aumenta, lo que indica que el test presentaría una mejor consistencia interna sin este ítem.

6. Estimación de la puntuación verdadera

Recordemos que según la teoría clásica la puntuación empírica del sujeto es igual a la puntuación verdadera más el error aleatorio de medida. Conocer la precisión con la que se mide el constructo permite calcular la cantidad de error que está afectando a la puntuación empírica. Así, el hecho de conocer la fiabilidad del instrumento permite estimar la puntuación verdadera del sujeto. No obstante, debe tenerse en cuenta que no se puede calcular exactamente la puntuación verdadera del sujeto, aunque sí estimarla a partir del cálculo de un intervalo de confianza.

Fundamentalmente se utilizan dos procedimientos para valorar esta puntuación verdadera:

- La estimación que asume la distribución normal del error aleatorio.
- La estimación a partir del modelo de regresión lineal.

6.1. Estimación de la puntuación verdadera a partir de la distribución normal del error aleatorio

Este procedimiento asume que el error se distribuye según la ley normal con media 0 y varianza S_e^2 . A partir del cálculo de un intervalo de confianza se obtienen los límites inferior y superior entre los que se encontrará la puntuación verdadera (V) del sujeto con un determinado nivel de confianza (o bien asumiendo un determinado riesgo alfa). Los pasos que habría que seguir son los siguientes:

1. Calcular el error típico de medida (S_e) que viene dado por la expresión siguiente:

$$S_e = S_x\sqrt{1 - r_{xx}}$$

S_x: Desviación típica de las puntuaciones del test.

r_{xx}: Coeficiente de fiabilidad obtenido.

2. Buscar el valor $Z_{\alpha/2}$ para el nivel de confianza con el que se quiera trabajar. Si se ha fijado el nivel de confianza al 95% (o bien el riesgo de error al 5%), el valor $Z_{\alpha/2}$ que corresponde según las tablas de la distribución normal es de 1,96.

3. Calcular el error máximo de medida ($E_{máx}$) que se está dispuesto a asumir:

$$E_{máx} = Z_{\alpha/2} \cdot S_e$$

4. Calcular el intervalo de confianza en el que se encontrará la puntuación verdadera del sujeto a partir de la expresión siguiente:

$$IC = X \pm E_{máx}$$

Ejemplo

Se ha administrado un test a una muestra de 300 sujetos. La puntuación total del test presenta una media de 15,6 puntos, una desviación típica de 5,4 y un coeficiente de fiabilidad de 0,77. Si se quiere trabajar con un nivel de confianza del 95%, ¿entre qué valores se situaría la puntuación verdadera de un sujeto que ha obtenido una puntuación empírica de 18?

Error típico de medida: $S_e = 5,4\sqrt{1 - 0,77} = 2,59$

Un nivel de confianza del 95% corresponde a una $z_{\alpha/2} = 1,96$

Error máximo:

$$E_{máx} = 1,96 \cdot 2,59 = 5,08$$

$$IC = 18 \pm 5,08$$

$$12,92 \leq V \leq 23,08$$

La puntuación verdadera del sujeto oscilará entre los valores 12,92 y 23,08, con un nivel de confianza del 95%.

Hay que tener en cuenta que la precisión de la medida, es decir, la fiabilidad del instrumento que utilizamos para medir el constructo, tiene un efecto directo en la amplitud del intervalo construido. Cuando utilizamos un instrumento con una fiabilidad alta, la estimación de la puntuación verdadera del sujeto será más precisa, lo que implica que la amplitud del intervalo será menor. Sucede lo contrario cuando se utilizan instrumentos menos fiables: la precisión con la cual podemos calcular la puntuación verdadera del sujeto es menor y el intervalo construido es más amplio.

6.2. Estimación de la puntuación verdadera a partir del modelo de regresión lineal

La puntuación empírica es una puntuación sesgada (teóricamente se observa una correlación positiva entre la puntuación empírica y el error de medida), por lo que la estimación que se pueda hacer a partir de esta también presentará sesgo. Por este motivo, otra posibilidad es primero valorar la puntuación verdadera y después calcular el intervalo de confianza a partir de la puntuación verdadera obtenida.

Si nos basamos en el supuesto de que la puntuación verdadera es igual a la media de la puntuación empírica, podemos hacer una estimación de la puntuación verdadera a partir de la ecuación de regresión siguiente:

$$V' = r_{xx}(X - \bar{X}) + \bar{X}$$

V': Puntuación verdadera pronosticada.

r_{xx}: Coeficiente de fiabilidad del test.

X: Puntuación empírica obtenida por el sujeto.

$\bar{X}$: Media de las puntuaciones del test.

Una vez estimada la puntuación verdadera, se pasaría a calcular el intervalo de confianza para establecer con un determinado nivel de confianza entre qué valores se situaría la puntuación verdadera del sujeto. En este caso se utilizará el

error típico de estimación (desviación típica de la diferencia entre la puntuación verdadera y la puntuación verdadera pronosticada ($V - V'$)):

$$S_{vx} = S_x\sqrt{1 - r_{xx}}\sqrt{r_{xx}} = S_e\sqrt{r_{xx}}$$

S_x: Desviación típica de las puntuaciones del test.

r_{xx}: Coeficiente de fiabilidad del test.

S_e: Error típico de medida.

Continuando con este ejemplo, a continuación se muestra cómo calcular primero la puntuación verdadera pronosticada y después el intervalo de confianza que indicará entre qué valores se situará la puntuación verdadera del sujeto con un nivel de confianza del 95%.

En primer lugar se calcula la puntuación verdadera:

$$V' = 0,77 \cdot (18 - 15,6) + 15,6 = 17,45$$

Se obtiene el error típico de estimación a partir del error típico de medida calculado en el ejemplo anterior y el coeficiente de fiabilidad:

$$S_{vx} = 2,59\sqrt{0,77} = 2,27$$

A continuación se calcula el error máximo:

$$E_{máx} = 1,96 \cdot 2,27 = 4,45$$

El intervalo de confianza quedará construido a partir de la estimación de la puntuación verdadera y el error máximo obtenido.

$$IC = 17,45 \pm 4,45$$

$$13,00 \leq V \leq 21,90$$

El resultado del intervalo, ya sea a partir de la estimación basada a partir de la distribución normal de los errores o a partir del modelo de regresión lineal, se interpreta del mismo modo. En los dos casos se puede afirmar que existe una probabilidad del 95% de que un sujeto que haya tenido una puntuación empírica de 18 puntos en el

test sitúe su puntuación verdadera entre 13 y 23 (o entre 13 y 22, si la estimación se ha realizado a partir del modelo de regresión lineal).

7. Fiabilidad de los tests referidos al criterio

7.1. Conceptos básicos

Los tests a los que se ha hecho referencia hasta el momento son los llamados tests referidos a la norma (TRN). Tal como se ha expuesto en el capítulo "Aproximación histórica y conceptos básicos de la psicometría", estos tests tienen como objetivo escalar los sujetos en función de las puntuaciones obtenidas en un test que mide un determinado rasgo o variable psicológica. En este apartado nos centraremos en otro tipo de tests, los tests referidos al criterio (TRC).

Estos tipos de tests empiezan a hacerse populares hacia la década de los años setenta del siglo xx, cuando consiguen hacerse un lugar dentro de la teoría de los tests para evaluar los estándares de rendimiento. Su surgimiento se explica, por un lado, por el contexto contrario que se había desarrollado, fundamentalmente en Estados Unidos sobre el uso de los tests, y por otro, por la necesidad de valorar la eficacia de los programas educativos que durante los últimos años habían surgido. En este último punto habría que evaluar el nivel de habilidad y conocimiento que los individuos alcanzaron dentro del programa educativo para demostrar su eficacia. En la actualidad, su uso está muy extendido en los campos educativo y laboral, donde lo que se persigue es evaluar la competencia que presentan los individuos en un determinado conjunto de conocimientos o criterio.

Un TRC pretende evaluar en términos absolutos el estatus que exhibe un sujeto respecto a un criterio, entendiendo como criterio un dominio de conductas, contenidos o conjunto de procesos muy definido. Para lograr este punto, primero es necesario definir el conjunto de tareas que el individuo debe ser capaz de realizar para demostrar su competencia sobre el criterio. El test se construye a partir de una serie de ítems que representan el dominio que quiere ser evaluado. La estimación del criterio se realiza a partir de la puntuación que el individuo obtiene en el test. La interpretación de esta puntuación es bastante intuitiva, dado que la

proporción de ítems correctamente acertados indicará la competencia que el sujeto tiene del criterio. Por ejemplo, si se pretende valorar la capacidad de matemáticas de los alumnos de tercero de primaria, habrá que definir el dominio de conductas (de contenidos o de procesos) que los alumnos deberían poder realizar. Una vez decididos los ítems (preguntas, problemas, ejercicios, etc.) que formarán parte del test, y que representan el criterio que se pretende valorar, estos se administran a los sujetos. La puntuación que se obtenga será la estimación del criterio.

En este punto vale la pena detenerse a valorar las diferencias entre los TRN y los TRC, dado que, como se mostrará en los apartados siguientes, en la práctica la distinción no está siempre clara. Mientras que los TRN tienen como finalidad saber si la puntuación del sujeto A es superior a la del sujeto B cuando se valora X, los TRC se plantean si los sujetos A y B son capaces de resolver X.

En la tabla siguiente se resumen las características básicas de estos dos tipos de test respecto al objeto que se evalúa, el análisis de la fiabilidad que realizan, el objetivo de la evaluación y el ámbito de aplicación.

Tabla 11. Características básicas de los tests referidos a la norma y de los tests referidos al criterio

	TRN	TRC
Objeto que se evalúa	Mide una variable psicològica o rasgo	Mide un conjunto de conocimientos o competencias (criterio)
Anàlisis de la fiabilidad	A partir de las diferencias entre los sujetos	A partir del dominio que presenta el individuo sobre el tema
Objetivo de evaluación	Posición relativa de un individuo respecto al resto del grupo	Grado de conocimiento que presentan los individuos en el dominio
Àmbito de aplicación	Personalidad, actitudes, etc.	Educación, ámbito laboral, evaluación de programas, etc.

En esencia, el concepto de fiabilidad en el caso de los TRC y los TRN es idéntico: en los dos casos se pretende valorar el grado de error que se comete a la hora de hacer una medición. No obstante, los procedimientos que se siguen para valorar la fiabilidad en la teoría clásica para los TRN no son, en general, apropiados para estimar la fiabilidad de los TRC. La razón es que mientras que

los TRN basan la medida de fiabilidad en la variabilidad de las puntuaciones del test (recordemos que el coeficiente de fiabilidad se ha definido como la proporción de varianza de las puntuaciones verdaderas que hay en la varianza de las puntuaciones empíricas), en los TRC esta variabilidad deja de tener importancia. La finalidad que persiguen los TRC es evaluar el nivel de conocimiento que los individuos poseen sobre un criterio, y mayoritariamente basan la medida de fiabilidad en clasificar a los sujetos de manera consistente en dos grandes categorías: aquellos que dominan el criterio y aquellos que no lo dominan. Si la clasificación que se hace de los individuos a partir de una o más administraciones del test es consistente, se puede hablar de consistencia o precisión en el proceso de medida y por lo tanto, de fiabilidad. No obstante, a diferencia de los TRN, los TRC se centran en la idea de que la puntuación del test permite hacer una interpretación en términos absolutos de la capacidad del sujeto sin tener que comparar los resultados con un grupo de referencia. Las diferencias entre los individuos (la variabilidad de la medida) dejan de tener importancia para evaluar la fiabilidad del test (las diferencias en la puntuación del test entre unos y otros pueden ser pequeñas, sin que esto afecte a la fiabilidad).

Como se ha visto antes, uno de los factores que afecta a los valores de la fiabilidad es la variabilidad de las puntuaciones del test; por eso los procedimientos que hasta ahora se han presentado para valorar la fiabilidad no son adecuados para ser aplicados a los TRC. Siguiendo a Shrock y Coscarelli (2007), hay tres maneras de abordar la fiabilidad en este tipo de test: aquellos procedimientos que requieren dos aplicaciones del test para valorar la consistencia de la clasificación, aquellos que solo requieren una única aplicación y aquellos en los que entra en juego el papel de los evaluadores. En este último caso, son los propios evaluadores los que juzgan la competencia de los sujetos. En los apartados siguientes se tratarán estos diferentes procedimientos para abordar la fiabilidad en los TRC.

7.2. Índices de acuerdo que requieren dos aplicaciones del test

Se considera que un test es fiable cuando se administra el mismo test en dos ocasiones a la misma muestra o cuando se aplican dos formas paralelas del test y los resultados de las dos administraciones permiten clasificar a los sujetos den-

tro de la misma categoría. A continuación se presentan dos de los índices de acuerdo más utilizados: el coeficiente de Hambleton y Novick ($p_{H\text{-}N}$) y el coeficiente kappa (k).

7.2.1. Coeficiente de Hambleton y Novick

El coeficiente de Hambleton y Novick (1973) valora la fiabilidad o consistencia de las clasificaciones a partir de dos administraciones del test o de dos formas paralelas del test. Para proceder con su cálculo se tiene en cuenta, por un lado, la proporción de sujetos que son clasificados de manera consistente en una misma categoría (pc), sea esta la de competentes o la de no competentes (aptas o no aptas, etc.), y por otro lado, la proporción de clasificaciones consistentes que se esperaría por azar (p_a) .La diferencia entre la proporción de valoraciones consistentes y aquellas que se esperarían por azar da el coeficiente de Hambleton y Novick ($p_{H\text{-}N}$):

$$p_{H-N} = p_c - p_a$$

El cálculo de la proporción de sujetos clasificados de modo consistente se expresa de la manera siguiente:

$$p_c = \sum p_i = \frac{n_{11}}{N} + \frac{n_{00}}{N}$$

p_c: Proporción de sujetos clasificados consistentemente en las dos administraciones del test.

N: Número total de sujetos evaluados.

n_{11}: Número de sujetos clasificados como competentes en las dos administraciones del test.

n_{00}: Número de sujetos clasificados como no competentes en las dos administraciones del test.

A pesar de que la expresión que se proporciona contempla clasificar a los individuos solo en dos categorías (competentes y no competentes), la fórmula se

puede aplicar independientemente del número de categorías. En este caso habrá que sumar las proporciones de cada una de las categorías establecidas.

Cuando todos los sujetos son clasificados de manera consistente en las dos administraciones del test, p_c toma el valor máximo de 1. Su valor mínimo viene determinado por la proporción de clasificaciones consistentes que se esperaría por azar (p_a). Para calcular esta proporción de clasificaciones consistentes que se esperarían por azar se utilizan las frecuencias marginales de una tabla de contingencia aplicando la fórmula siguiente:

$$p_a = \sum \frac{n_{.j} \cdot n_{i.}}{N^2}$$

$n_{.j}$: Número de sujetos clasificados como competentes (o no competentes) por el test A.

$n_{i.}$: Número de sujetos clasificados como competentes (o no competentes) por el test B.

Ejemplo

Supongamos que se administran dos tests paralelos de 20 ítems a un grupo de 16 individuos. Para que el individuo sea clasificado en el grupo de sujetos competentes se requiere que conteste correctamente 14 ítems.

En la tabla siguiente se muestran los datos de este ejemplo. En esta tabla se representan las puntuaciones obtenidas por los 16 individuos en los dos tests (A y B) y la clasificación que a partir del punto de corte establece cada uno de estos tests (clasificación test A y clasificación test B).

Tabla 12. Datos de ejemplo. Coeficiente de Hambleton y Novick

Sujeto	Test A	Test B	Clasificación test A	Clasificación test B
1	10	12	No competente	No competente
2	16	18	Competente	Competente
3	19	20	Competente	Competente
4	8	10	No competente	No competente
5	10	12	No competente	No competente

Sujeto	Test A	Test B	Clasificación test A	Clasificación test B
6	15	16	Competente	Competente
7	14	12	Competente	No competente
8	17	18	Competente	Competente
9	18	19	Competente	Competente
10	13	14	No competente	Competente
11	16	17	Competente	Competente
12	19	17	Competente	Competente
13	12	15	No competente	Competente
14	10	12	No competente	No competente
15	16	14	Competente	Competente
16	18	15	Competente	Competente

A partir de la tabla anterior, se construye la tabla de contingencia, que veremos a continuación.

Tabla 13. Tabla de contingencia. Coeficiente de Hambleton y Novick

		Test B		
Test A		Competentes	No competentes	
	Competentes	9 (n_{11})	1	10 (n_{i1})
	No competentes	2	4 (n_{00})	6 (n_{i2})
		11(n_{1j})	5 (n_{2j})	16 (N)

A partir de la tabla de contingencia podemos calcular el coeficiente de Hambleton y Novic:

$$p_c = \frac{9}{16} + \frac{4}{16} = 0{,}81$$

$$p_a = \frac{10 \cdot 11}{16^2} + \frac{6 \cdot 5}{16^2} = 0,546 \approx 0,55$$

$$p_{H-N} = 0,81 - 0,55 = 0,26$$

Este resultado indica que el uso de los tests permite mejorar un 26% la clasificación de los sujetos de la que se realizaría por azar.

7.2.2. Coeficiente kappa de Cohen

El coeficiente kappa (Cohen, 1960) permite estudiar el nivel de concordancia en las clasificaciones a partir de dos administraciones del test. Posiblemente este sea el coeficiente de consistencia más extensamente utilizado en la literatura. Su fórmula viene dada por la expresión siguiente:

$$k = \frac{p_c - p_a}{1 - p_a}$$

Donde p_c y p_a son, respectivamente, la proporción de sujetos clasificados de manera consistente y la que se esperaría por azar tal como se ha definido antes. Valores cercanos a 1 indican que la consistencia en la clasificación de los sujetos a partir del test es perfecta, mientras que valores cercanos a 0 indican que la consistencia en la clasificación es debida al azar (en este caso la aplicación de los tests no ha mejorado la consistencia que por azar se podría obtener). En general, los valores del coeficiente kappa que oscilan entre 0,6 y 0,8 se consideran aceptables y aquellos que se sitúan por encima de 0,8 se interpretan como muy buenos (Landis y Koch, 1977).

A partir del uso del error típico de medida (S_e), propuesto también por Cohen (1960), puede obtenerse el intervalo de confianza y valorar su significación estadística. El error típico de medida y el intervalo de confianza vienen definidos a partir de las expresiones siguientes:

$$S_{e(k)} = \sqrt{\frac{p_c(1 - p_c)}{N(1 - p_a)^2}}$$

$$IC = k \pm Z_{\alpha/2} \cdot S_{e(k)}$$

A partir de los datos expuestos en el ejemplo anterior el resultado del coeficiente kappa sería:

$$k = \frac{0,81 - 0,55}{1 - 0,55} = 0,58$$

Este valor indica la consistencia en la clasificación de los sujetos independientemente de la proporción esperada por el azar. A partir del error típico de medida se obtiene el intervalo de confianza con un nivel de confianza del 95%:

$$S_{e(k)} = \sqrt{\frac{0,81 \cdot (1 - 0,81)}{16 \cdot (1 - 0,55)^2}} = 0,22$$

$$IC_{95\%} = 0,58 \pm 1,96 \cdot 0,22$$

$$0,15 \leq k \leq 1$$

Hay que señalar que debido al reducido tamaño de muestra el intervalo construido es excesivamente amplio. Si en vez de tener 16 sujetos fueran 200 los que se hubieran evaluado, se conseguiría un intervalo bastante más preciso:

$$S_{e(k)} = \sqrt{\frac{0,81 \cdot (1 - 0,81)}{200 \cdot (1 - 0,55)^2}} = 0,06$$

$$IC_{95\%} = 0,58 \pm 1,96 \cdot 0,06$$

$$0,46 \leq k \leq 0,70$$

7.2.3. Coeficiente de Livingston

Los dos procedimientos anteriores consideran que un error en la clasificación de los individuos es igual de grave, independientemente de que la puntuación del sujeto se sitúe cerca o lejos respecto al punto de corte. No obstante, la lógica nos dice que si un individuo ha obtenido una puntuación muy distanciada del punto de corte (es decir, muestra una competencia sobradamente alta o muy inferior a la del punto de corte), sería difícil que a partir de una segunda aplicación del test o de un test paralelo el resultado de la clasificación fuera contrario al obtenido en la primera administración. En este sentido, Livingston (1972) propone un nuevo coeficiente para evaluar la fiabilidad de las clasifica-

ciones en el que tiene en cuenta este aspecto. Este nuevo coeficiente se basa en los métodos de pérdida cuadrática que tienen en cuenta la distancia que existe entre el punto de corte y la media de las puntuaciones del test. El coeficiente de Livingston se expresa a partir de la fórmula siguiente:

$$K^2_{xx'} = \frac{r_{xx} S_x S_{x'} + (\overline{x}_x - C)(\overline{x}_{x'} - C)}{\sqrt{[S^2_x + (\overline{x}_x - C)^2][S^2_{x'} + (\overline{x}_{x'} - C)^2]}}$$

Lectura de la fórmula:

r_{xx}: Coeficiente de fiabilidad a partir del procedimiento de formas paralelas o test-retests.

S_x y $S_{x'}$: Corresponden, respectivamente, a la desviación típica del test en la primera y segunda administración o en cada una de las formas paralelas del test.

$\overline{x}_x$ y $\overline{x}_{x'}$: Corresponden, respectivamente, a la media del test en la primera y segunda administración o en cada una de las formas paralelas del test.

C: Punto de corte.

S^2_x y $S^2_{x'}$: Corresponden, respectivamente, a la varianza del test en la primera y segunda administración o en cada una de las formas paralelas del test.

A continuación se muestra la aplicación de la fórmula en una situación práctica.

Supongamos que se han diseñado dos formas paralelas de un examen de psicometría. Al aplicarlas a una muestra de estudiantes se obtuvo que la media y desviación típica del test A fue de 5,2 y 2,6, respectivamente. Mientras que en el test B se obtuvo una media de 5,4 y una desviación típica de 2,3. La correlación entre los dos tests fue de 0,83. El punto de corte para determinar el aprobado en la prueba se situó en el 5,5. A partir de estos datos podemos calcular $K^2_{xx'}$:

$$K^2_{xx'} = \frac{0,83 \cdot 2,6 \cdot 2,3 + (5,2 - 5,5) \cdot (5,4 - 5,5)}{\sqrt{[2,6^2 + (5,2 - 5,5)^2] \cdot [2,3^2 + (5,4 - 5,5)^2]}} = 0,83$$

Este resultado muestra que la concordancia de clasificación a partir de los dos tests es buena.

7.3. Índices de acuerdo que requieren una única aplicación del test

La desventaja principal de los indicadores presentados en el apartado anterior es que hay que aplicar el test dos veces o generar una forma paralela del test. A raíz de este hecho surgió la necesidad de encontrar algún procedimiento en el que solo se necesitará una única administración del test.

La primera propuesta para evaluar la fiabilidad de los TRC que solo requiriera una única administración del test vino de la mano de Livingston (1972), quien propuso una leve modificación a la formulación que se ha presentado en el apartado anterior. A partir de esta primera propuesta siguieron otras muchas[2]. La mayoría de estos procedimientos basan el cálculo de la consistencia de la clasificación utilizando modelos estadísticos que estiman la puntuación que se habría obtenido en una segunda administración del test a partir de la puntuación empírica obtenida por los sujetos. No obstante, hay que decir que la mayoría de estos procedimientos son complejos y poco utilizados en la práctica profesional. Por ello, en este apartado solo se abordará la propuesta inicial de Livingston (1972) por ser una de las más sencillas de cálculo y posiblemente la más utilizada en la práctica (Shrock y Coscarelli, 2007).

7.3.1. Coeficiente de Livingston (una única aplicación)

La propuesta que se ha presentado del coeficiente de Livingston en el apartado anterior puede ser fácilmente modificada para que sea válida cuando solo se cuenta con una única administración del test. En este caso, se debe tener en cuenta como coeficiente de fiabilidad el coeficiente de consistencia interna del test. La fórmula propuesta por Livingston (1972) viene dada por la siguiente expresión:

$$K^2 = \frac{r_{xx} S_x^2 + (\overline{x} - C)^2}{S_x^2 + (\overline{x} - C)^2}$$

2. Subkoviak (1976), Huynh (1976), Brennan y Kanne (1977), Breyer y Lewis (1994), Livingston y Lewis (1995), Brennan y Wan (2004), Lee (2005), Lee, Brennan, Wan (2009), entre otros.

r_{xx}: Coeficiente de fiabilidad a partir del procedimiento de las dos mitades, KR_{20} o alfa de Cronbach.

S_x^2: Varianza de las puntuaciones del test.

$\overline{x}$: Media de las puntuaciones del test.

C: Punto de corte.

Si solo contáramos con los resultados del primer examen de psicometría del ejemplo anterior y el coeficiente alfa de Cronbach fuera de 0,78, K^2 sería:

$$K^2 = \frac{0,78 \cdot 2,6^2 + (5,2 - 5,5)^2}{2,6^2 + (5,2 - 5,5)^2} = 0,783$$

Hay que tener presente que cuando la media coincide con el punto de corte, K^2 será igual al coeficiente de fiabilidad. Por ejemplo, en el caso anterior se puede observar que la distancia entre la media y el punto de corte no es muy elevada, por lo que el valor de K^2 es muy similar al valor del coeficiente alfa de Cronbach. En cambio, se observa que cuando el punto de corte se distancia de la media, K^2 aumenta. Si ahora consideráramos que el punto de corte es de 6,5, el resultado sería:

$$K^2 = \frac{0,78 \cdot 2,6^2 + (5,2 - 6,5)^2}{2,6^2 + (5,2 - 6,5)^2} = 0,82$$

Asimismo, se observa que cuando aumenta el coeficiente de fiabilidad del test también aumenta K^2. Ahora, si aumentásemos el coeficiente de alfa de Cronbach a 0,85, observaríamos que también aumentaría el resultado de K^2:

$$K^2 = \frac{0,85 \cdot 2,6^2 + (5,2 - 5,5)^2}{2,6^2 + (5,2 - 5,5)^2} = 0,852$$

Por ello se demuestra que K^2 siempre será igual o mayor que el coeficiente de fiabilidad del test, y que cuando el coeficiente de fiabilidad sea 1, K^2 tomará también el valor máximo de 1.

$$K^2 \geq r_{xx}$$

7.4. Fiabilidad interobservadores

Los procedimientos presentados en los apartados previos hacían referencia a la capacidad del test para poder clasificar de manera consistente a los sujetos. En este apartado se presentarán algunos de los indicadores más frecuentes para poder evaluar la consistencia de las evaluaciones realizadas por diferentes jueces. En algunos contextos educativos y de empresa, es frecuente que la competencia de un individuo pueda ser evaluada por diferentes observadores. En estos casos la calidad de la medida depende de la consistencia que se observa entre la evaluación de los observadores. El coeficiente kappa para datos nominales u ordinales y el coeficiente de concordancia para variables continuas son los dos coeficientes más ampliamente utilizados de correlación que se presentarán en este apartado[3].

7.4.1. Coeficiente kappa

En este caso el coeficiente kappa se aplica cuando se quiere estudiar la concordancia que existe entre las valoraciones realizadas por dos evaluadores. La fórmula, tal como se ha presentado antes, viene dada por la expresión siguiente:

$$k = \frac{p_c - p_a}{1 - p_a}$$

Donde p_c corresponde a la proporción de sujetos clasificados de manera consistente por los dos (o más) evaluadores y p_a corresponde a la proporción de concordancias que se esperaría que sucedieran entre los dos evaluadores por azar.

Supongamos que dos evaluadores han clasificado en dos categorías a 80 trabajadores que han realizado un curso de formación: aquellos que han logrado las competencias básicas para poder desarrollar las nuevas tareas (competentes)

3. Recordad que hemos visto el coeficiente kappa en el apartado "Coeficiente kappa de Cohen" en este mismo capítulo.

y aquellos que se considera que no (no competentes). En la tabla de contingencia siguiente se presentan los resultados de sus valoraciones:

Tabla 14

	Observador 1			
Observador 2		Competentes	No competentes	
	Competentes	49 (n_{11})	11	60 (n_{i1})
	No competentes	1	19 (n_{00})	20 (n_{i2})
		50 (n_{1j})	30 (n_{2j})	80(N)

$$p_c = \frac{n_{11}}{N} + \frac{n_{00}}{N} = \frac{49}{80} + \frac{19}{80} = 0,85$$

$$p_a = \sum \frac{n_{.j} \cdot n_{i.}}{N^2} = \frac{60 \cdot 50}{80^2} + \frac{30 \cdot 20}{80^2} = 0,523$$

$$k = \frac{p_c - p_a}{1 - p_a} = \frac{0,85 - 0,523}{1 - 0,523} = 0,69$$

Este valor se interpretaría como una consistencia aceptable entre los dos evaluadores.

7.4.2. Coeficiente de concordancia

Lin (1989) propuso un coeficiente de fiabilidad para calcular el grado de acuerdo en las valoraciones realizadas por dos evaluadores cuando estas fueran de carácter continuo. A pesar de que en apartados previos se ha visto que cuando las variables son continuas se aplica el coeficiente de correlación de Pearson (r_{xy}), en este caso no sería adecuado aplicarlo. La razón es que el coeficiente de correlación de Pearson valora solo si el orden de las valoraciones de los dos evaluadores coinciden, pero no si los valores asignados a estas valoraciones son realmente los mismos. Por ello, se podrían obtener coeficientes de correlación

de Pearson altos (si la ordenación entre los dos evaluadores es similar), a pesar de que ninguna valoración fuera coincidente. El coeficiente de concordancia propuesto por Lin (1989) permite valorar el grado en el que los valores absolutos otorgados por cada evaluador son concordantes. El coeficiente viene definido por la expresión siguiente:

$$CC = \frac{2r_{xy}S_xS_y}{S_x^2 + S_y^2 + (\overline{x} - \overline{y})^2}$$

r_{xy}: Valor del coeficiente de correlación de Pearson entre las dos valoraciones.

S_x y S_y: Corresponden respectivamente a la desviación típica del evaluador 1 y del evaluador 2.

S_x^2 y S_y^2: Corresponden respectivamente a la varianza del evaluador 1 y del evaluador 2.

$\overline{x}$ y $\overline{y}$: Corresponden respectivamente a la media de las evaluaciones realizadas por el evaluador 1 y 2.

El coeficiente puede tomar valores entre 1 y –1, pero por la cuestión que interesa valorar (coincidencia de las puntuaciones entre los evaluadores) solo tendrá sentido cuando este tome valores positivos. Por otro lado, hay que tener en cuenta que a la hora de valorar el resultado, este coeficiente interpreta el grado de acuerdo de manera mucho más exigente: valores superiores a 0,99 se interpretan como una concordancia casi perfecta; entre 0,95 y 0,99 se habla de una concordancia sustancial; entre 0,90 y 0,95, moderada, y por debajo de 0,90 se considera que la concordancia es pobre.

Ejemplo

Dos evaluadores corrigen de manera independiente los trabajos de 40 alumnos. La correlación de Pearson entre los dos evaluadores es de 0,96. La media de las calificaciones del evaluador A es de 7,23, con una desviación típica de 3,44, mientras que para el evaluador B la media es de 6,33, con desviación típica de 2,88. ¿Cuál es el coeficiente de concordancia entre los dos evaluadores?

$$CC = \frac{2 \cdot 0,96 \cdot 3,44 \cdot 2,88}{3,44^2 + 2,88^2 + (7,23 - 6,33)^2} = 0,91$$

El coeficiente de concordancia indica que el grado de acuerdo entre los dos evaluadores es moderado.

8. Estimación de los puntos de corte

Los procedimientos que se han presentado hasta ahora requieren previamente establecer un punto de corte para calcular la fiabilidad. Tal como apunta Berk (1986), el punto de corte es el punto que permite tomar decisiones y clasificar a los sujetos como competentes (aquellos que dominan el criterio) y no competentes (aquellos que no lo dominan). Dado que el punto de corte es una puntuación empírica del test, tiene asociado un error como puntuación del test que es y un error como puntuación a partir de la cual se toman decisiones. El primer error es el error de medida, que se ha presentado en los apartados previos y que se expresa a partir del error típico de medida. El segundo error es un error de clasificación y puede tomar dos formatos:

- Un individuo que es competente y que erróneamente es clasificado como no competente (falso no competente).
- Un individuo que no es competente y que erróneamente es clasificado como competente (falso competente).

A pesar de que es importante evitar estos dos tipos de errores, hay situaciones en las que cometer un tipo de error u otro tiene repercusiones diferentes y representa una mayor o menor gravedad.

Supongamos que después de un curso de formación en una empresa se evalúa la competencia de los trabajadores para poder llevar a cabo una nueva tarea. Una vez finalizado el curso, los participantes realizan un examen para valorar si han logrado las competencias básicas para enfrentarse a la nueva tarea. La empresa decide que aquellos que no han superado la prueba participen en unas nuevas charlas para intentar

lograr esta competencia. En una situación de este tipo se considera que las consecuencias de clasificar a un sujeto como competente cuando no lo es resultan más graves que clasificar a un sujeto como incompetente cuando en realidad no lo es. En el primer caso, el trabajador no sabrá enfrentarse a la nueva tarea, se equivocará al realizarla, con la posibilidad de que su mala praxis provoque otras consecuencias más graves. En el segundo caso se envía a un sujeto competente a recibir unas charlas adicionales, que seguramente le permitirán lograr con más firmeza los conocimientos que ya tenía adquiridos.

En este apartado abordaremos algunos de los métodos más frecuentes para establecer este punto de corte. Tal como sugieren Cizek y Bunch (2007), cada uno de los métodos mezcla una parte de arte y una parte de ciencia. Por otro lado, hay que tener en cuenta que cada uno de los métodos puede llevar a un resultado diferente (diferente punto de corte y diferente porcentaje de sujetos clasificados en cada grupo) y que en definitiva, a pesar de que algunos métodos son más adecuados en función del tipo de tests o circunstancia, ninguno ha demostrado ser superior a los otros. Todos tienen en común que unos jueces expertos determinarán el punto de corte de manera sistemática a partir de la evidencia que tienen sobre aquello que pretenden valorar. Según Jaeger (1989), los diferentes métodos pueden ser clasificados en dos grandes grupos: por un lado, los métodos que se centran en la valoración del test (*test centered*) y, por otro, aquellos que se centran en la ejecución de los examinandos (*examinee centered*). No obstante, por motivos didácticos la exposición que se seguirá de los diferentes métodos se realizará según tres criterios:

- Métodos que se basan en la valoración que un grupo de expertos o jueces realizan sobre los ítems de un test (*test centered*). Se presentarán los métodos de Nedelsky (1954), Angoff (1971, 1984) y Sireci, Hambleton y Pitoniak (2004).
- Métodos que se basan en la valoración de un grupo de expertos sobre la competencia de los sujetos (*examinee centered)*. Aquí se expondrán los métodos del grupo de contraste (Berk, 1976) y del grupo límite (Zieky y Livingston, 1977).
- Métodos que se basan en la posición del sujeto respecto al grupo normativo, o también llamados métodos de compromiso, que intentan ligar aspectos referentes a los TRC y a los TRN. En este apartado se presentarán los métodos de Hofstee (1983) y Beuk (1984).

8.1. Métodos basados en la evaluación de expertos sobre los ítems

Los métodos que se presentan a continuación tienen en común que los participantes serán jueces expertos sobre el contenido de la materia que evalúan. En todos los casos se les requerirá que valoren si los ítems del test pueden ser contestados correctamente por determinados individuos, sin que por ello sea necesario tener información real sobre la competencia de los individuos a los que se administrará la prueba. De este último aspecto se desprende una de las ventajas principales de estos métodos, y es que pueden ser aplicados antes de que los sujetos contesten el test, dado que no se necesitan datos sobre su ejecución.

8.1.1. Método de Nedelsky

El método que propuso Nedelsky (1954) se aplica sobre ítems de respuesta múltiple. Se utiliza todavía hoy en día de manera amplia en el ámbito académico para determinar el punto de corte y poder determinar si los individuos tienen los conocimientos mínimos sobre una materia concreta. Un concepto clave de este procedimiento y otros que se tratarán en este apartado es que los expertos o jueces deben tener en mente un hipotético grupo de sujetos que se encontrarían en el límite para considerarlos competentes (o incompetentes).

El método se basa en que un grupo de expertos, sobre el contenido que se pretende valorar, determina para cada ítem las alternativas de respuesta que un sujeto con los conocimientos mínimos requeridos sobre la materia para superar la prueba (con la competencia mínima requerida) rechazaría como incorrectas. El recíproco de las alternativas restantes (1/número de alternativas restantes) es lo que se denomina *valor Nedelsky*. Este valor se interpreta como la probabilidad de que un individuo con una capacidad mínima sobre la materia seleccione la alternativa correcta.

Para ilustrar el procedimiento, imaginemos un ítem con cinco alternativas de respuesta. El juez determina que dos de las alternativas de respuesta pueden ser descartadas fácilmente por un sujeto que presente los conocimientos mínimos requeridos sobre la materia. El sujeto, por lo tanto, se enfrenta a tres alternativas restantes. El valor Nedelsky en este caso será 1/3 = 0,33. Este procedimiento habría que repetirlo para cada ítem del test, y posteriormente sumar todos los valores obtenidos para cada uno de los ítems. La suma de todos los valores se utiliza como punto de corte del test, que servirá para clasificar a los

individuos en las categorías de competentes y no competentes. No obstante, hay que tener en cuenta que es habitual que en el proceso participe más de un juez o evaluador, por lo que habrá que hacer la media de las valoraciones de cada uno de los jueces y posteriormente hacer el sumatorio de las valoraciones medias. En la tabla siguiente se muestra un ejemplo de cálculo del punto de corte a partir de la valoración de 4 jueces a un examen de 12 ítems con 5 alternativas de respuesta.

Tabla 15. Ejemplo de cálculo del método Nedelsky

	Evaluadores				
	A	**B**	**C**	**D**	
Ítems	**Valores Nedelsky**				**Medias**
1	0,5	0,33	0,5	0,5	0,4575
2	1	0,5	1	1	0,875
3	0,2	0,25	0,2	0,25	0,225
4	0,33	0,5	0,25	0,25	0,3325
5	0,33	0,33	0,33	0,25	0,31
6	0,2	0,33	0,5	0,5	0,3825
7	1	0,33	0,5	0,33	0,54
8	0,5	0,25	0,33	0,5	0,395
9	0,25	0,2	0,2	0,2	0,2125
10	0,33	0,33	0,5	0,5	0,415
11	1	1	0,5	0,5	0,75
12	0,33	0,5	0,5	0,5	0,4575
Sumatorio					**5,3525**

En este ejemplo el sumatorio de todas las medias es de 5,35, lo que indica que un sujeto con la competencia mínima requerida sobre la materia para superar la prueba debería contestar 5 ítems correctamente. Así pues, se recomendaría que los sujetos con 5 ítems correctos sean clasificados como competentes, es decir, que aprueben el examen.

La principal limitación del método es que los valores que pueden adquirir las probabilidades asignadas a los ítems son muy limitados y no presentan interva-

los iguales entre ellos. Por ejemplo, en el caso anterior los valores de Nedelsky pueden ser: 0,2, 0,25, 0,33, 0,5 y 1, y como se puede observar las distancias entre ellos no son iguales. Este hecho lleva a que, en los casos en los que se observa mayor distancia entre los valores de la probabilidad (por ejemplo entre los valores 0,5 y 1), muchos evaluadores tiendan a puntuar preferiblemente los ítems con una probabilidad de 0,5 antes que con una probabilidad de 1 (probabilidad de que un sujeto con la capacidad mínima requerida sobre el contenido sea capaz de descartar todas las alternativas incorrectas). Esta actuación (otorgar una probabilidad de 0,5 en vez de 1) implica un sesgo a la baja en la puntuación de corte de la prueba, lo que provoca que este método proporcione puntos de corte más bajos en comparación a otros métodos.

8.1.2. Método de Angoff

El método de Angoff (1971) es el más utilizado en la práctica y el que presenta más variantes respecto a su planteamiento inicial (Cizek y Bunch, 2007). Como el método de Nedelsky, consiste en que un grupo de evaluadores base sus juicios teniendo en mente un hipotético grupo de individuos que presenta una competencia mínima sobre la materia para poder superar una prueba. Hay que destacar que, a diferencia del método plateado por Nedelsky, Angoff propuso que los jueces se fijaran en el ítem globalmente y no en cada una de las alternativas de respuesta.

La propuesta inicial de Angoff (1971) consistía en que el evaluador debía determinar si un individuo hipotético, con la competencia mínima requerida, sería capaz de contestar el ítem correctamente. Al ítem se le asignaba un 1, si se considera que el sujeto es capaz de contestar correctamente el ítem, y se puntuaba con un cero si se consideraba que esta persona fallaría el ítem. El sumatorio de las puntuaciones asignadas a cada ítem sería igual a su punto de corte de la prueba, lo que permitiría distinguir aquellos sujetos competentes de los que no presentaran una competencia mínima requerida sobre la materia. No obstante, la versión más extendida de este método requiere la que el propio Angoff planteó en un pie de página en su descripción inicial del método. En este caso proponía una variación del método que consistía en que en lugar de pensar en un individuo concreto, los jueces hicieran sus valoraciones teniendo en mente

a un grupo de individuos con una competencia mínima para superar la prueba y que valoraran la proporción de sujetos que contestarían el ítem correctamente. El sumatorio de las probabilidades asignadas a cada uno de los ítems representaría el punto de corte de la prueba. El ejemplo siguiente muestra un hipotético caso de esta última opción.

Se solicita a cuatro evaluadores que indiquen el porcentaje de sujetos que sería capaz de resolver correctamente cada uno de los 12 ítems de un examen de psicometría. Se indica a los evaluadores que hay que centrar sus valoraciones pensando en un posible grupo de individuos que supuestamente presenta la competencia mínima requerida sobre la materia. En la tabla siguiente se muestran los resultados de este hipotético caso.

Tabla 16. Ejemplo de cálculo del método de Angoff

	Evaluadores			
	A	**B**	**C**	**D**
Ítems	**Porcentaje de sujetos que resuelven correctamente el ítem**			
1	30	40	45	38
2	80	90	75	85
3	85	70	65	70
4	50	45	50	60
5	60	50	45	55
6	80	90	70	75
7	40	45	50	35
8	20	25	30	10
9	90	80	85	90
10	30	40	35	40
11	40	45	50	50
12	60	65	70	75
Media	55,42	57,08	55,83	56,92

El punto de corte en este caso se obtendrá a partir del cálculo de la media de las puntuaciones otorgadas por los cuatro jueces:

$$C = \frac{55,42 + 57,08 + 55,83 + 56,92}{4} = 56,31\%$$

Este resultado indica que el punto de corte recomendado por este test sería que los sujetos contestaran correctamente un 56,31% de los ítems, es decir, 6,76 ítems (7 ítems) del conjunto de 12 ítems del test.

Una ventaja importante de este método es que el hecho de evaluar globalmente el ítem permite utilizar ítems de otros formatos, además de los ítems de elección múltiple utilizados en el método de Nedelsky. Esto supone que el método puede ser utilizado cuando los ítems de la prueba son en formato abierto (Hambleton y Plake, 1995). Como se ha comentado antes, las variantes del método son muchas. Una muy usada para evitar opiniones divergentes entre los jueces es pedir a los evaluadores que valoren los ítems en diferentes rondas, empleando un procedimiento similar al que se utiliza en un método Delphi.

8.1.3. Método del consenso directo

El método del consenso directo es un método relativamente reciente propuesto por Sireci, Hambleton y Pitoniak (2004). Algunas de las ventajas destacables respecto a los métodos anteriores son que, por un lado, cada juez experto que participa en el proceso puede expresar directamente su opinión sobre cuál debería ser la localización concreta del punto de corte y, por otro, que permite modificar a los jueces sus valoraciones según las opiniones ofrecidas por otros participantes.

El procedimiento que se sigue en este método requiere que los ítems del test se agrupan en secciones. Estas secciones están formadas a partir de subáreas de contenido homogéneas. La función de los jueces consiste en decir cuántos ítems de cada sección podrían ser contestados correctamente por sujetos que tienen una competencia mínima para poder ser considerados competentes. El sumatorio de los ítems en cada subárea que se considera que se contestarán correctamente será igual a la puntuación de corte.

Inmediatamente después de que los jueces han realizado sus valoraciones, se presentan a todo el panel de expertos los resultados de cada uno de los miembros y se procede a discutir las razones de las posibles diferencias. En esta fase se promueve una discusión abierta entre los jueces para que defiendan y razonen el porqué de sus valoraciones. El objetivo es que la discusión facilite el consenso entre los evaluadores que permita una convergencia en el punto de corte. Una vez finalizada la discusión, se les ofrece la oportunidad de volver a valorar cada una de las secciones del test. Por último, una vez recogidos los datos de la segunda valoración, se comentan en una segunda ronda las diferentes puntuaciones para intentar una mayor convergencia en el punto de corte del test.

Se solicitó a cuatro jueces que valoraran una prueba de psicometría de 60 ítems. La prueba se dividió en cuatro secciones (análisis de los ítems, fiabilidad, validez y transformación de puntuaciones) y cada sección presentaba un número diferente de ítems. En la tabla siguiente se presenta por columnas el número de ítems que cada juez considera que contestará correctamente un sujeto con conocimientos mínimos para superar la prueba. En la última fila se recoge el sumatorio de ítems de cada experto, que representa su punto de corte recomendado para la prueba. En las dos últimas columnas se muestran las medias y desviaciones típicas (*DT*) de estas valoraciones y el porcentaje de ítems que serían contestados correctamente en cada sección.

Tabla 17. Ejemplo hipotético del método del consenso directo

	Evaluadores					
	A	**B**	**C**	**D**		
Secciones	**Número de ítems con-testados correctamente**				**Media y DT**	**%**
Anàlisis de los ítems (14 ítems)	8	7	8	8	7,75 (0,5)	64,58
	Evaluadores					
	A	**B**	**C**	**D**		
Secciones	**Número de ítems con-testados correctamente**				**Media y DT**	**%**
Fiabilidad (20 ítems)	14	12	13	10	12,25 (1,71)	68,06
Validez (16 ítems)	10	11	10	11	10,50 (0,58)	65,63
Transformación de puntuaciones (10 ítems)	6	7	8	7	7,00 (0,82)	70
Sumatorio	38	37	39	36	37,5	62,50

Las desviaciones típicas indican la variabilidad de opinión entre los jueces para cada sección. En el ejemplo, los jueces tienen juicios bastante homogéneos al valorar los ítems referentes a validez (a pesar de que los ítems que hayan podido señalar cada uno de los jueces puedan ser en cada caso diferentes), dado que la desviación típica respecto a su media es menor. En cambio, la sección de fiabilidad presenta un mayor desacuerdo entre los jueces, dado que su desviación típica respecto a la media es la más alta en relación con el resto de las secciones. El objetivo de la discusión de estos resultados con el panel de jueces es que se promueva una mayor convergencia en los datos, que ayude a lograr un consenso en el punto de corte que habría que tomar en esta prueba.

8.2. Métodos basados en la evaluación de expertos sobre la competencia de los sujetos

Los métodos que se presentarán en este apartado se caracterizan por que los jueces, además de ser expertos en la materia que evalúan, han de conocer la competencia de los sujetos. Este hecho implica que las valoraciones que realizarán no serán sobre sujetos hipotéticos, sino sobre individuos reales. Básicamente es en este aspecto donde radica la crítica más feroz sobre estos métodos, dado que a la hora de hacer las valoraciones, los evaluadores pueden estar influenciados no solo por los conocimientos o habilidades de los sujetos, sino por otras variables intrínsecas al individuo que poco tienen que ver con la competencia, como por ejemplo el sexo, la personalidad, la raza, el comportamiento, etc. Otros tipos de sesgo que se han observado a la hora de hacer evaluaciones por parte de jueces se pueden consultar en la revisión de Martínez-Arias (2010).

8.2.1. Método del grupo de contraste

El método del grupo de contraste fue propuesto inicialmente por Berk (1976). En este método es necesario que los jueces clasifiquen a los sujetos en dos grupos en función del nivel de competencia que suponen que exhibirán en la materia evaluada. Es necesario pues que los jueces conozcan sobradamente el rendimiento que exhibirán los sujetos que deben clasificar. En un grupo clasifican a los individuos que consideran que serán competentes y en el otro a los que consideran

que serán no competentes. Una vez realizada esta clasificación hay que administrar la prueba a los sujetos y puntuarla. Para determinar el punto de corte, se suelen utilizar diferentes procedimientos: representar gráficamente las puntuaciones de la prueba de cada uno de los grupos, utilizar como punto de corte algún indicador de tendencia central (media o mediana) o elaborar el análisis de regresión logística. No obstante, el uso de este último procedimiento solo se recomienda si las muestras son suficientemente grandes (Cizek y Bunch, 2007).

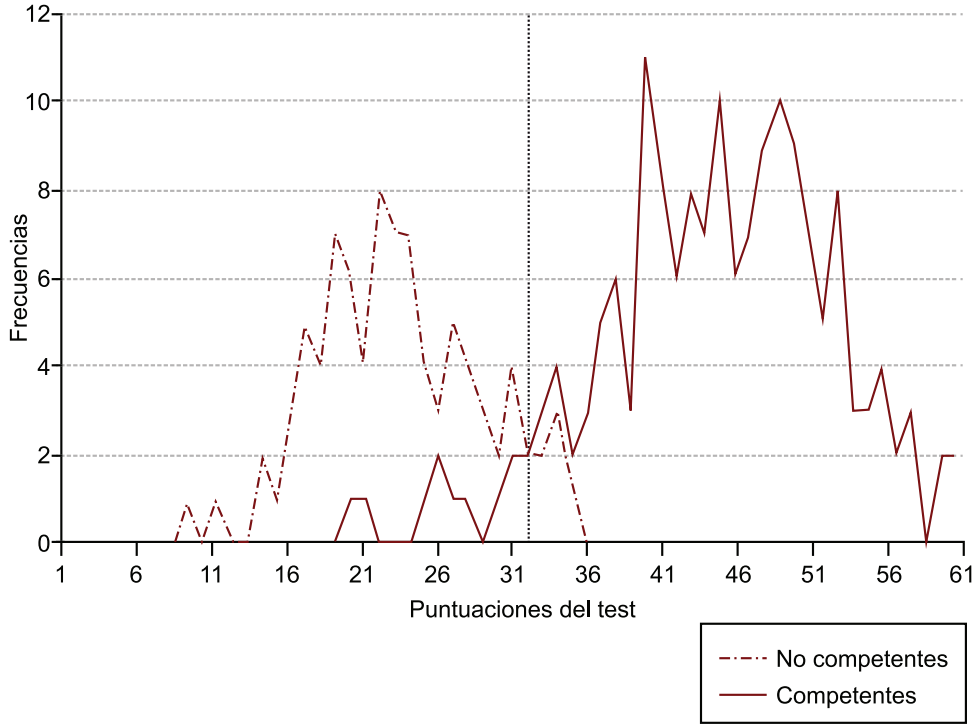
Posiblemente, determinar el punto de corte a partir de la representación gráfica sea el método más sencillo. Consiste en representar en una misma gráfica las distribuciones de los grupos: los que han sido clasificados como competentes por los jueces y los que han sido valorados como no competentes (figura 2). La intersección entre ambas distribuciones sería el punto de corte de la prueba. Idealmente se busca aquel punto de corte en el que todos los sujetos valorados como competentes se sitúen por encima del punto de corte, mientras que los no competentes se encuentren por debajo. En la práctica, tal como se ve en la figura 2, hay un solapamiento entre las dos distribuciones, con lo que en función de dónde se sitúe el punto de corte habrá más sujetos competentes clasificados que no competentes o al revés.

Por ejemplo, si observamos la figura 2, si se moviera la línea vertical (que indicaría el punto de corte) hacia la izquierda, el número de falsos negativos se reduciría, es decir, se disminuiría la posibilidad de considerar un competente como no competente (a pesar de que se aumentaría el número de sujetos no competentes, que se consideran competentes). Mientras que si la línea se moviera hacia la derecha, disminuirían los falsos positivos, es decir, aquellos que siendo no competentes se los considera competentes. Así pues, para fijar el punto de corte, hará falta en cada caso valorar qué tipo de error se quiere evitar.

Buscar el punto de corte a partir de la figura 2 puede resultar una tarea bastante subjetiva, por lo que algunos autores prefieren basar la posición del punto de corte a partir de la media o la mediana de los dos grupos.

A partir de los datos que hemos representado en la figura 2, a modo de ejemplo, supongamos que un grupo de profesores clasificó a sus alumnos de la asignatura de *Psicometría* en competentes y no competentes (n = 258). Los profesores clasificaron a 90 alumnos como no competentes y a 168 como competentes. Los estudiantes hicieron la prueba y la mediana del grupo de no competentes fue de 22 puntos y la del grupo de competentes de 44. El punto medio entre las dos medianas (en este caso 33) puede funcionar como punto de corte de la prueba.

Figura 2. Método del grupo de contraste. Distribuciones del grupo de competentes y no competentes en las puntuaciones de un test



8.2.2. Método del grupo límite

En muchas ocasiones puede resultar difícil a los jueces clasificar a los sujetos claramente en dos grupos, como competentes o no competentes. En este sentido, el método del grupo límite (Zieky y Livingston, 1977) viene a paliar este aspecto y puede ser utilizado como alternativa al método de los grupos de contraste. En este método, generalmente se pide a los jueces que clasifiquen a los sujetos en tres grupos: un grupo de sujetos que claramente son competentes, otro grupo que claramente se percibe como no competentes y finalmente otro que se situaría entre los dos grupos anteriores. Es decir, un grupo límite, en el que los sujetos tendrían una ejecución que los situaría en medio del grupo de competentes y no competentes. Una vez se ha clasificado a los sujetos se administra la prueba. Habitualmente, el valor de la mediana del grupo límite en la

prueba es el que se utiliza como punto de corte. Dada la sencillez del método, otros autores (Plake y Hambleton, 2001) propusimos una versión alternativa para valorar grupos de sujetos con competencia básica, notable y avanzada.

8.3. Métodos de compromiso

El término *método de compromiso* fue propuesto por Hofstee (1983) para recoger la idea de que se necesitaban nuevos procedimientos que combinaran, por un lado, los conocimientos que un sujeto presentaba sobre la materia (valoración de la competencia en términos absolutos), pero también el nivel de ejecución que presentaba respecto a su grupo normativo (valoración de la competencia en términos relativos). Los métodos que se han presentado hasta ahora para fijar el punto de corte son métodos que se basan exclusivamente en la competencia que el sujeto exhibe sobre la materia que debe evaluar (valoración en términos absolutos). No obstante, en la práctica los juicios que emiten los evaluadores no pueden estar estrictamente basados en un criterio absoluto, sino que en cualquier proceso de evaluación se tiene en cuenta información referente al grupo normativo de referencia.

Un ejemplo muy clarificador que ilustra el uso combinado de información sobre el criterio y sobre el grupo normativo en cualquier proceso de evaluación lo proporcionan Cizek y Bunch (2007). Estos autores proponen una situación en la que se valora la competencia de control de esfínteres de un niño. Los padres siguen las rutinas habituales para que su niño adquiera un control adecuado, hasta que llega el gran día en el que pueden afirmar que su hijo ha adquirido la competencia. Hasta aquí todo parece bastante corriente, pero ¿cambiaría la percepción de la situación si dijéramos que el niño de este ejemplo tiene 9 años? Evidentemente, sí.

Este ejemplo demuestra que las personas utilizan múltiples fuentes para valorar las diferentes situaciones a las que se enfrentan y que cualquier intento de anular o reducir alguna de estas fuentes ocasionaría una valoración artificial y poco ajustada a la realidad.

Los métodos que se presentan a continuación aceptan que los tests referidos a la norma y al criterio están, en realidad, bastante unidos a la hora de valorar a los individuos, dado que utilizan las normas para poder fijar el criterio. En este apartado se presentarán dos de los procedimientos más habituales: el método de Hofstee (1983) y el método de Beuk (1984).

8.3.1. Método de Hofstee

Hofstee propuso este método en 1983, cuando después de impartir la misma asignatura durante algunos años se encontró que en aquel curso los alumnos presentaban un rendimiento muy bajo respecto a años anteriores (sin que ningún aspecto significativo respecto al material, profesores, etc., variaran de un año al otro). Después de rebajar la nota de corte a un 4,5, solo el 55% de los estudiantes aprobaban la asignatura (cuando en años anteriores la nota de corte se había situado en un 6 y aun así aprobaba el 90% de los alumnos). El método que Hofstee propone requiere que los evaluadores respondan a cuatro preguntas sobre los sujetos que serán evaluados:

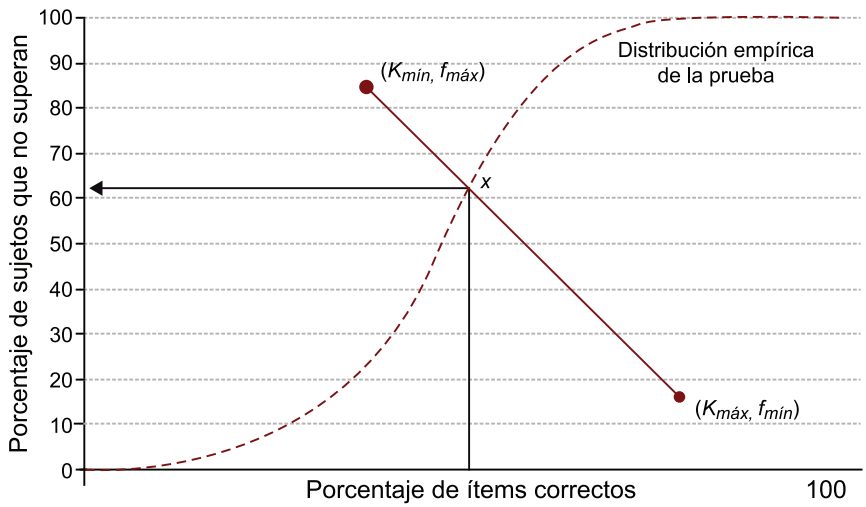
1) ¿Cuál es el punto de corte más alto que se considera aceptable, a pesar de que todo el mundo llegara a esta puntuación? Se simboliza por $k_{máx}$.
2) ¿Cuál es el punto de corte más bajo que se considera aceptable, a pesar de que nadie llegara a esta puntuación? Se simboliza por $k_{mín}$.
3) ¿Cuál es el porcentaje máximo de sujetos que se toleraría que no superaran la prueba? Se simboliza por $f_{máx}$.
4) ¿Cuál es el porcentaje mínimo de sujetos que se toleraría que no superaran la prueba? Se simboliza por $f_{mín}$.

Como se observa, dos de las preguntas se basan en el nivel de conocimiento que los evaluados deben presentar ($k_{máx}$ y $k_{mín}$). Estas dos cuestiones se miden a partir del porcentaje de ítems que hay que contestar correctamente. Las otras dos preguntas se basan en el porcentaje de no competentes que se toleraría dada una determinada prueba de evaluación ($f_{máx}$ y $f_{mín}$). A partir de estos cuatro puntos y la distribución empírica de la prueba, se encuentra el punto de corte (x) óptimo para este método de compromiso.

En la figura 3 se representa gráficamente cómo encontrar este punto de corte. Por un lado, en el eje de abscisas se representa el porcentaje de ítems correctos de la prueba y, por otro, en el eje de ordenadas el porcentaje de personas que no superan la prueba. En esta gráfica también se representa la distribución empírica de la prueba, que muestra que a medida que aumenta el número de ítems que hay que contestar correctamente (eje de abscisas), aumenta el porcentaje de personas que no superan la prueba. Los valores k y f se representan con dos puntos en el eje de

coordenadas y se unen con una línea recta que cruza la distribución empírica de la prueba (en realidad se espera que en la mayoría de las ocasiones atravesará la distribución empírica). El punto en el que se cruza la distribución empírica y la recta *k-f* indicará el porcentaje de ítems correctos que habrá que exigir a la prueba (el punto de corte) y el correspondiente porcentaje de sujetos que no superan la prueba si se utiliza este punto de corte. Este punto de corte se simboliza en la figura con *x*.

Figura 3. Representación gráfica del método de Hofstee



8.3.2. Método de Beuk

El método propuesto por Beuk (1984) surgió como una simplificación del método de Hofstee (1983). Como el anterior, presupone que los evaluadores tienen una idea más o menos clara sobre cuál es el punto de corte que sería necesario aplicar a la prueba y cuál debería ser el porcentaje de personas que tendrían que superarla. En este caso, se pide a los evaluadores que respondan a las siguientes dos preguntas:

1) ¿Cuál debería ser el porcentaje mínimo de ítems que habría que contestar correctamente para superar la prueba? Este valor se simboliza con *x*.

2) ¿Cuál es el porcentaje de personas que se espera que superen la prueba? Este valor se simboliza con *y*.

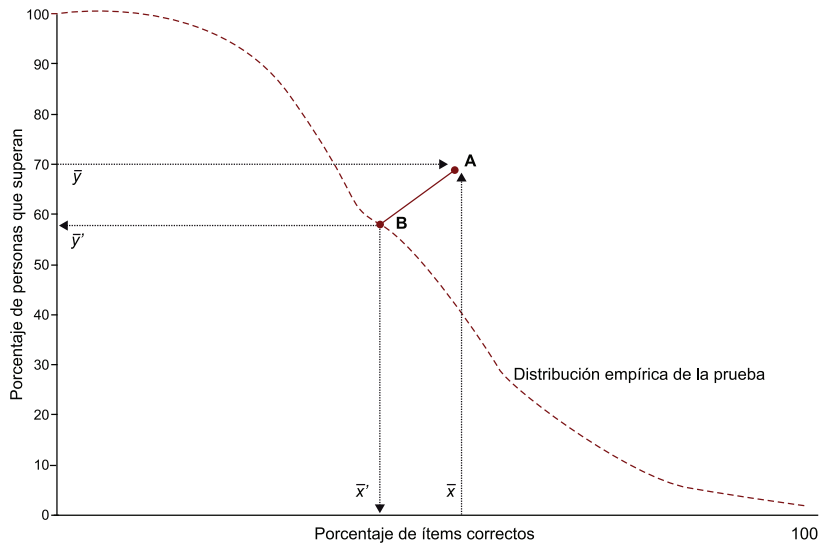
A partir del cálculo de la media de las valoraciones de los evaluadores a estas dos preguntas ($\bar{x}$ e $\bar{y}$) y la distribución empírica obtenida de la aplicación de la prueba, se obtiene el punto de corte. El ejemplo que se presenta a continuación y la figura 4 muestran cómo obtener el punto de corte a partir de este método.

Supongamos que como media los evaluadores han determinado que haría falta que un 60% de los ítems de la prueba se contestaran correctamente $\bar{x}$ y que sería necesario que un 70% de las personas que realizaran la prueba la superaran $\bar{y}$.

En la figura 4 se representan estos dos puntos y su intersección se simboliza con la letra A. Se representa la distribución empírica de la prueba (línea de puntos), que en este caso será decreciente, dado que el número de personas que superarán la prueba (eje de ordenadas) disminuirá a medida que el porcentaje de ítems que hay que contestar de manera correcta (eje de abscisas) aumente.

El paso siguiente es calcular las desviaciones típicas a las dos preguntas formuladas a los evaluadores. El cociente entre las dos desviaciones típicas (S_x/S_y) será la pendiente de la recta que atravesará la distribución empírica de la prueba. El punto en el que la recta atraviesa la distribución empírica se simboliza con el punto B. Finalmente, como se puede ver en la figura 4, la proyección de este punto en el eje de abscisas proporcionará el punto de corte de la prueba (porcentaje de ítems correctos) y su proyección sobre el eje de ordenadas proporcionará el porcentaje de sujetos que superarán la prueba.

Figura 4. Representación gráfica del método de Beuk



Capítulo III

Validez

Luis Manuel Lozano
Jaume Turbany

Cuando un psicólogo decide aplicar un cuestionario es para alcanzar un objetivo determinado. Para ello, debe asegurarse de que el cuestionario que va a usar posee unas adecuadas propiedades psicométricas. Entre ellas cabe destacar la que vamos a tratar en este capítulo: la validez.

La validez es uno de los aspectos más importantes, quizá el que más, tanto en la elaboración como en la evaluación de cuestionarios psicológicos. A fin de cuentas, se trata de comprobar que la utilización del test está siendo correcta y que los objetivos que desea alcanzar el psicólogo que lo utiliza son factibles.

En el apartado "¿Qué es la validez?" se hace un breve recorrido histórico sobre dicho concepto. Como se puede observar, es un concepto que ha estado evolucionando (y aún lo está) hasta que se alcanza la idea que actualmente está en vigor. Dicho concepto se define oficialmente en los estándares publicados en 1999 conjuntamente por la American Educational Research Methods (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME). Estas entidades defienden que se pueden agrupar los indicios de validez de un test en cinco apartados: evidencia basada en la validez de contenido, basada en el proceso de respuesta, basada en la estructura interna del cuestionario, basada en la relación con otras variables y basada en las consecuencias de la evaluación.

En los siguientes apartados se trata cada uno de los indicios de validez previamente mencionados y se buscan estrategias para poder obtener dichos indicios (a excepción del apartado de las consecuencias de la evaluación, en el que solo se tratan las consecuencias que se pueden esperar de la aplicación de un cuestionario).

En el último apartado, "Factores que afectan a la validez", se tratan diferentes aspectos que afectan a algunas de las técnicas expuestas con anterioridad para determinar los diferentes indicios de validez.

1. Qué es la validez

1.1. Definición

Para comprender el concepto de validez es necesario realizar un pequeño estudio de la evolución histórica que ha sufrido dicho concepto.

La utilización de cuestionarios se vio impulsada por la primera y segunda guerras mundiales. Durante esos momentos se tuvo la necesidad de incorporar al ejército a la población civil, destinándola al puesto más adecuado. Tras rellenar los cuestionarios se comprobaba en el campo de entrenamiento si los sujetos rendían satisfactoriamente o no en el puesto al que se les había destinado. Dado que en primer lugar se hacía la medición y posteriormente se evaluaba el éxito, se hablaba de validez predictiva. Es decir, un test posee *validez predictiva* si sirve para predecir el comportamiento en un constructo que será evaluado posteriormente a la aplicación del cuestionario.

Más tarde se trató de evaluar la relación existente entre las características de las personas que realizaban un trabajo y su éxito en él. De este modo, se trataba de conocer qué características podrían predecir el éxito laboral y buscarlas cuando se realizaba una selección de personal. Dado que el estudio se realizaba sobre personas que ya tenían el puesto y se valoraba su ejecución, se hablaba de validez concurrente, ya que ambas mediciones se hacían a la vez. Es decir, un test posee *validez concurrente* si sirve para predecir el comportamiento en un constructo que es evaluado simultáneamente a la aplicación del cuestionario.

Como se puede observar, inicialmente los tests eran exclusivamente empleados para predecir. Así pues, en un comienzo, se consideraba que un test era válido si servía para predecir alguna variable de interés, denominada *criterio* (Guilford, 1946).

Por lo tanto, se conceptualiza la validez como correlación entre el cuestionario y el criterio de interés (ya sea evaluado con posterioridad o simultáneamente a la aplicación del cuestionario). Así pues, se considera que un test es válido para evaluar cualquier aspecto con el que correlacione (Bingham, 1937; Guilford, 1946; entre otros).

Uno de los problemas de la conceptualización de la validez como correlación es el hecho de que hay que encontrar una medida del criterio adecuada, es decir, se necesitan datos del criterio que hayan sido obtenidos de una manera fiable y válida. Por tanto, si ya se posee una medida válida del criterio, ¿para qué se necesita aplicar un cuestionario?

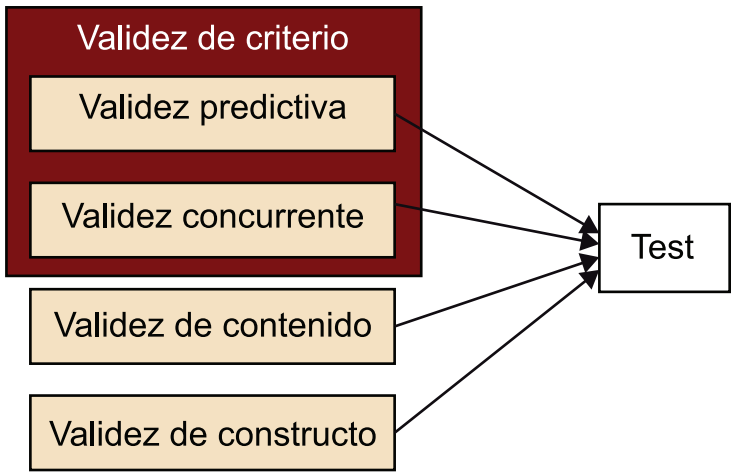
Otro problema de esta conceptualización es que dejaba fuera a un gran número de tests educativos. En estos no se trata de predecir la conducta, se trata de comprobar cuánto se ha aprendido después de un periodo de formación. En estos cuestionarios la puntuación obtenida es un indicador de lo que el test pretende evaluar (conocimiento en matemáticas, en inglés, etc.) y no un predictor de criterios distintos del test. Desde esta perspectiva, la validez hace referencia a que los ítems que componen el cuestionario sean representativos de aquello que se pretende evaluar. A este concepto se le denominó *validez de contenido* (Anastasi, 1954).

Por otro lado, a lo largo de los años treinta se produce un auge de las teorías que tratan de conocer la estructura factorial de la inteligencia. Con estas teorías comienza a conceptualizarse un test como válido cuando representa de manera fidedigna el constructo psicológico que pretende medir, así como las relaciones esperadas entre los diferentes constructos. De este modo nace la validez de constructo (Cronbach y Meehl, 1955). Las técnicas estadísticas empleadas para poder comprobar dicha validez son, tradicionalmente, el análisis factorial exploratorio y las matrices multirrasgo-multimétodo (Campbell y Fiske, 1959), y más recientemente el análisis factorial confirmatorio. Por ejemplo, si se emplea un test que evalúa la triada cognitiva desde el modelo cognitivo de depresión de Beck (Beck, Rush, Shawn y Emery, 1979) (pensamientos sobre mí mismo, pensamientos sobre el mundo y pensamientos sobre el futuro), el cuestionario tendrá *validez de constructo* si evalúa las tres dimensiones y estas tienen las relaciones que se esperan con, por ejemplo, ansiedad.

Hasta los años ochenta se podía hablar de validez predictiva, validez concurrente, validez de contenido y validez de constructo de un cuestionario, si bien

las dos primeras en los estándares de los tests y manuales educativos y psicológicos publicados por la APA, AERA y NCME en 1966 y 1974 se englobaban como *validez de criterio*.

Figura 1



Posteriormente, Cronbach (1971) puntualizó que en un test que pretende medir un rasgo de personalidad no existe solo un criterio relevante que predecir, ni un contenido que muestrear (validez predictiva y de contenido respectivamente). Se dispone, por el contrario, de una teoría acerca del rasgo y de sus relaciones con otros constructos y variables (validez de constructo). Si se hipotetiza que la puntuación del test es una manifestación válida del atributo, se puede contrastar la asunción analizando sus relaciones con otras variables. Por tanto, comenzó a existir una tendencia en la que consideraban la validez como algo unitario, siendo la validez de constructo la científicamente más admisible y estando la validez de criterio y de contenido incluidas en esta (Messick, 1989). Así pues, se impone la concepción de que la validación de constructo constituye un marco integral para obtener pruebas de la validez incluyendo las procedentes de la validación de criterio y de contenido. De hecho, deja de hablarse de las diferentes categorías de validez para comenzar a hablar de diferentes evidencias implicadas en los tres tipos tradicionales de validez (criterio, contenido y constructo).

Dado que tanto el estudio de la estructura del constructo como las relaciones de este con otros constructos pasa a ser considerado la principal forma de validez, este

proceso puede concebirse como un caso particular de la contrastación de las teorías científicas mediante el método hipotético-deductivo (Prieto y Delgado, 2010).

Notad que en estos momentos, a mediados de los años ochenta, existe un cambio muy relevante: mientras que al comienzo se conceptualiza la validez como una propiedad inherente al test, después se pasa a concebir que lo que realmente se valida no es el test en sí, sino las inferencias que se realizan a partir de este. Por ello, el responsable de asegurar la validez ya no es solo el constructor del test, sino que también lo es el usuario que emplea dicho cuestionario para una finalidad determinada. En muchas ocasiones los problemas que un cuestionario posea en lo referente a la validez se deben no al diseño del cuestionario sino a la utilización que se hace de este.

Actualmente, en la última edición hasta el momento de los *Standards for educational and psychological testing* (AERA, APA y NCME, 1999), muy influenciados por el capítulo escrito por Messick (1989) y el libro de Shepard, Camilli, Linn y Bohrnstedt (1993), se defiende que la validez hace referencia al *grado en el que la evidencia empírica y la teoría apoyan la interpretación de las puntuaciones de los tests relacionada con un uso específico*. Como se puede apreciar, se concibe la validez como un concepto unitario. Para comprobar la validez se debe atender a cinco evidencias de esta:

- *El contenido de test*: Los ítems que constituyen el test son relevantes y representativos del constructo psicológico que se desea medir.
- *El proceso de respuesta*: El proceso que siguen las personas al contestar al test permite extraer respuestas indicadoras de lo que se quiere evaluar.
- *La estructura interna*: Las relaciones de los ítems entre sí son congruentes con el modelo teórico empleado a la hora de definir el constructo que evaluar.
- *La relación con otras variables*: Las relaciones que se establecen entre el constructo que se evalúa y otros constructos son las esperadas según el marco teórico en el que se haya definido el constructo que evaluar.
- *Las consecuencias de la aplicación del cuestionario*: Las consecuencias tanto positivas como negativas que se extraen al emplear un test son las previstas.

Como breve resumen de lo expuesto anteriormente se presenta la siguiente tabla, en la que se puede apreciar la evolución del concepto en los diferentes estándares publicados por la APA.

Tabla 1

Edición	Validez
1954	Constructo, concurrente, predictiva, contenido
1966	Criterio, constructo, contenido
1974	Criterio, constructo, contenido
1985	Unitaria (pero mantienen criterio, constructo y contenido)
1999	Unitaria: 5 fuentes de evidencia

1.2. Importancia de la validez

El concepto de validez es central en psicometría. Tal y como se comentó anteriormente, para comprobar la validez se deben acumular evidencias que proporcionen una base científica para interpretar las puntuaciones de un cuestionario de manera adecuada. Por ello, lo que realmente se valida no es el cuestionario en sí, sino las interpretaciones que se hacen a partir de él. Por tanto, no se puede defender que un test sea válido o que por el contrario carezca de validez. Un test puede ser adecuado para un propósito pero no para otro.

Si se aplica un cuestionario con el que se pretende medir autoestima, las respuestas pueden ser empleadas con diferentes fines (conocer el nivel de autoestima de una persona para saber si es un problema que tratar en terapia, en selección de personal, como investigación sobre el propio constructo, etc.). Para poder usar el cuestionario con una finalidad determinada, se deben acumular evidencias que indiquen que el uso es correcto ("evidencias de validez"). En caso contrario, se estaría haciendo un mal uso de los tests, principales herramientas en el trabajo psicológico, y las conclusiones que se extrajeran de ellos no serían correctas. En el ejemplo anterior no se sabría si es un aspecto sobre el que se debe intervenir terapéuticamente, no se sabría si la persona seleccionada realmente tiene la autoestima que se desea o no se sabe si realmente se está midiendo autoestima.

Para poder realizar correctamente el trabajo como psicólogos, se debe saber si las conclusiones que se extraen a partir de los tests empleados son adecuadas, ya que en caso contrario se corre el riesgo de no saber exactamente qué se está evaluando o si esa medición realmente es útil para el propósito del psicólogo.

2. Evidencia de validez basada en el contenido

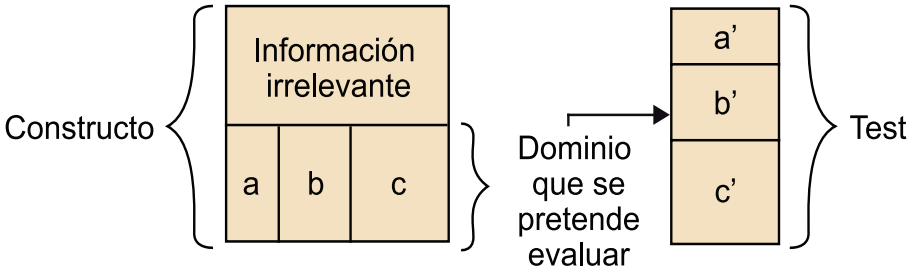
2.1. Concepto

Muchas de las inferencias y asunciones que se derivan de la interpretación de las puntuaciones en un test son más fácilmente evaluables si se examinan los procedimientos empleados para generar las puntuaciones. Por ejemplo, si se quiere inferir a partir de las puntuaciones en un test sobre determinada conducta o constructo psicológico, es de esperar que los ítems que componen el cuestionario sean tanto *relevantes* (que la información que se pregunta esté directamente relacionada con lo que se pretende medir), como *representativos* (las cuestiones que se realicen deben ser una muestra adecuada de todo lo que se pretende medir) de la conducta (Kane, 2006).

La evidencia de la validez de contenido hace referencia a la relación que existe entre los ítems que componen el test y lo que se pretende evaluar con él, prestando atención tanto a la relevancia como a la representatividad de los ítems. Este tipo de evidencia se recoge principalmente en el momento de la elaboración del test.

Supongamos que se desea elaborar un test para evaluar la personalidad. En este caso, se decide trabajar dentro del marco teórico de los cinco factores de la personalidad (extraversión, apertura, responsabilidad, amabilidad y neuroticismo). Dado que se trata de un test que se va a emplear en una selección de personal concreta, solo interesan las dimensiones de responsabilidad (a), amabilidad (b) y neuroticismo (c). En este ejemplo el constructo es la personalidad que está compuesta por las cinco dimensiones. Las dos primeras, para los intereses del test que se está realizando, son información irrelevante. Las otras tres son el dominio que interesa evaluar. A partir de este dominio se construyen ítems destinados a evaluar la responsabilidad (a'), la amabilidad (b') y el neuroticismo (c'). Dichos ítems deben tener relación con el factor que pretenden medir, es decir, los ítems que evalúan responsabilidad están relacionados con la definición que existe en la comunidad científica sobre dicho factor (relevancia). Pero a su vez los ítems deben preguntar por la totalidad del dominio que evaluar (representatividad).

Figura 2



En las pruebas educativas, las evidencias de validez basada en el contenido son fundamentales. Si no se comprueba que el test es consistente con los objetivos curriculares perseguidos (relevancia), es decir, que está libre de material irrelevante y que el que está representa adecuadamente el dominio que se pretende evaluar (representatividad), la utilidad del test se verá seriamente afectada y, por tanto, las conclusiones que se obtengan serán erróneas. En estas situaciones se suele recomendar, dado que el dominio que se quiere evaluar está claramente definido, emplear los diferentes métodos estadísticos de muestreo para obtener una muestra representativa de los contenidos que deben constituir el test (Muñiz, 2003).

El problema surge cuando no se dispone del dominio tan claramente definido. Por ejemplo, si se quiere realizar un test que evalúe la inteligencia, lo primero que se debe preguntar el constructor del cuestionario es: ¿qué es la conducta inteligente? En este caso, dado que no existe un dominio perfectamente definido, se deben buscar otras estrategias para obtener el indicador de la validez de contenido.

2.2. Procedimiento

En este apartado se presentará el procedimiento más habitual en la valoración de la evidencia basada en el contenido, si bien existen otros métodos menos empleados. Una recopilación de ellos se puede encontrar en Sireci (1998).

Si se quiere desarrollar un test, lo primero que se debe realizar es definir de manera operativa el dominio que evaluar. Tras realizar o aceptar una definición ya existente, se debe elaborar una *tabla de especificaciones*. Se trata de realizar una descripción detallada del test, determinar la proporción o el número de ítems

que evaluarán cada contenido o habilidad del dominio que evaluar; el formato de los ítems y de las respuestas (AERA, APA y NCME, 1999) (usualmente en este paso también se determinan las propiedades psicométricas que se desea que tenga la prueba).

Tras realizar los ítems se debe acudir a un grupo de expertos en la materia, que harán las veces de jueces. Para evitar cualquier sesgo, dichos jueces no deben estar implicados en la elaboración del cuestionario. Estos deben analizar cada uno de los ítems valorando en qué medida son *representativos* y *relevantes* para evaluar el dominio de interés, tomando como definición de este la aportada por los autores del test.

Se puede defender que existen, por tanto, tres aspectos bien diferenciados que se deben tener en cuenta a la hora de comprobar las evidencias de la validez de contenido: la definición del dominio, la representación de los ítems que evalúan el dominio y su relevancia (Sireci, 1998).

Es recomendable que la valoración de los ítems la realice cada juez por separado para, de este modo, evitar posibles sesgos a la hora de responder. Una vez que se poseen las valoraciones de todos los expertos, se deben buscar aquellos ítems en los que haya concordancia, seleccionándolos para formar parte del cuestionario.

Por ejemplo, si 8 de los 10 jueces determinan que un ítem destinado a medir depresión realmente evalúa lo que pretende, dicho ítem tendrá un índice de congruencia de 0,8. Se suelen considerar adecuados aquellos ítems que poseen un índice de congruencia igual o superior a 0,7 (Sireci, 1998).

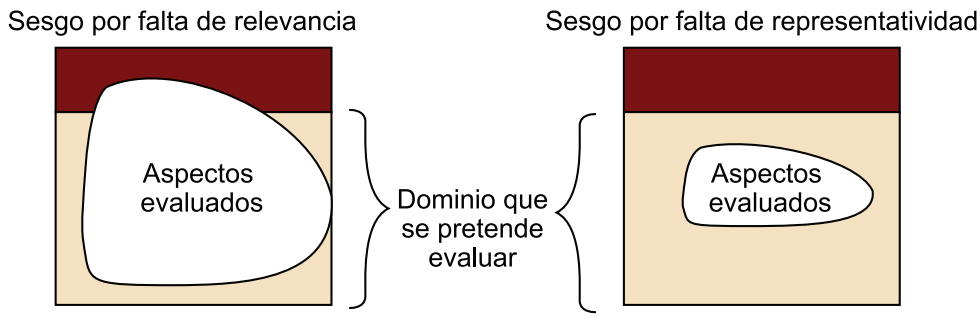
Los ítems en los que no haya acuerdo (que no alcancen un índice de congruencia de 0,7) no tienen por qué ser eliminados. Es recomendable que con estos ítems se realice un grupo de discusión con los expertos para que comenten las diferencias tratando de llegar a un punto de acuerdo para mejorar dichos ítems.

Este es el procedimiento más habitual a la hora de valorar los indicios de validez de contenido, si bien no está libre de críticas. El principal problema que se plantea en la utilización de expertos es que estos son altamente competentes en el contenido que se evalúa, por lo que pueden pasar por alto un texto cuyo nivel no sea adecuado para la comprensión de los sujetos que hay que evaluar o que puede ser fácilmente malinterpretado. Es decir, aunque el experto nos puede proporcionar información muy relevante, lo que realmente importa es cómo percibe y reacciona ante el test o el ítem la persona que está respondiendo (Leighton, 2004).

2.3. Contenido sesgado

El uso de expertos para valorar tanto la relevancia como la representatividad de los ítems tiene como finalidad evitar que el cuestionario tenga contenidos sesgados. Se dice que el contenido de un test está sesgado si los ítems que lo componen evalúan aspectos no relevantes para el dominio (*sesgo por falta de relevancia*) o si no representan de manera adecuada todo el dominio que se pretende evaluar (*sesgo por falta de representatividad*). Como se puede comprobar, un test está sesgado si no cubre adecuadamente el dominio que pretende medir o si incluye cuestiones no necesarias para valorar correctamente el dominio.

Figura 3

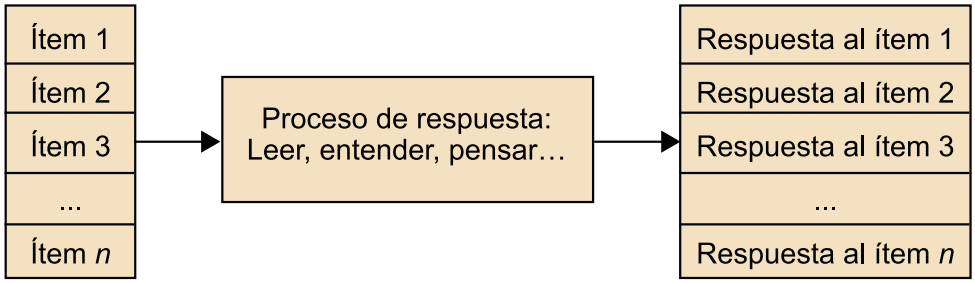


3. Evidencia de validez basada en el proceso de respuesta

3.1. Concepto

Este tipo de evidencia es un concepto que se introdujo como novedoso en los estándares publicados en 1999, si bien había sido previamente mencionado por algunos especialistas en la medida de lo psicológico, como Messick (1989). Los estándares describen este tipo de indicios como el ajuste entre el constructo evaluado y el proceso de respuesta realizada por las personas que responden al test.

Figura 4



Por *proceso de respuesta* se entienden todas las conductas que se necesitan para poder contestar un ítem, como pueden ser leer las preguntas, comprenderlas, decidir la respuesta que se quiere dar y finalmente responder al ítem.

Un ejemplo sobre este indicio de validez se puede encontrar en un examen de matemáticas que se realice a niños que están aprendiendo a sumar. Si el enunciado del ítem es: "3 + 2 = ", probablemente, si han adquirido el conocimiento necesario, pueden dar la respuesta correcta. Dicho ítem también puede tener un enunciado como: "Calcule el valor resultante de realizar una operación aditiva entre los valores 3 y 2". Evidentemente, un niño que esté aprendiendo a sumar puede responder al primer enunciado pero no al segundo (no tiene el vocabulario necesario, su capacidad lectora no le permitirá comprender la pregunta, etc.). Como resultado del primer enunciado se concluirá que ya tiene adquirido cierto nivel del dominio evaluado, pero con el segundo se concluirá que no. Es decir, dado que el segundo enunciado carece de validez de respuesta, llevará a los evaluadores a conclusiones erróneas sobre el nivel de habilidad del niño a la hora de realizar sumas.

A la hora de responder a un test se deben combinar las características de las preguntas realizadas, las de las respuestas que se pueden dar y las de la persona que responde. Por ello, existen diferentes factores que pueden afectar a la respuesta:

a) Factores relacionados con los ítems. En este apartado se deben tener en cuenta varios factores.

- Contenido de los ítems. Se debe asegurar que el contenido es adecuado a la población que se quiere evaluar. Por ejemplo, no se puede usar un test para evaluar depresión infantil si este fue construido para hacerlo

en adultos. Si se hace, se pueden encontrar preguntas del tipo "¿Ha sentido cambios en su deseo sexual?", que evidentemente son inapropiadas para evaluar a niños.

– Redacción de los ítems. El lenguaje empleado para realizar el ítem no debe superar la capacidad comprensiva de las personas que van a responder. Un ejemplo de esto puede observarse en el ejemplo de la suma expuesto anteriormente.
– Validez aparente del ítem. Cuando se evalúan conocimientos es deseable que las personas que responden al cuestionario piensen que es adecuado. Si se evalúa el conocimiento en psicometría mediante un test, es de esperar que los alumnos que rellenen el test piensen que las preguntas que se les realizan son adecuadas para medir el conocimiento en psicometría. Esto es lo que se denomina validez aparente. Por el contrario, en los tests de personalidad hay que tratar de que la persona que responde no sepa exactamente lo que se evalúa. De este modo se intenta evitar que responda lo que más le favorece o lo que piensa que se espera de él.

b) Factores relacionados con la respuesta a los ítems.

– El número de alternativas que se ofrezcan como respuesta. Los tests de actitudes suelen responderse en un formato tipo Likert. En estas escalas se les pide a las personas en qué grado están de acuerdo con la afirmación que se les presenta teniendo que responder en una escala donde, por ejemplo, 0 significa totalmente en desacuerdo y 5, totalmente de acuerdo. El problema surge cuando se emplea una escala que supera la capacidad discriminativa de las personas. En estudiantes universitarios una escala de 0 a 10 es perfectamente comprensible pero, por el contrario, esa misma escala empleada en personas sin estudios puede ser excesiva. Un universitario comprende perfectamente la diferencia entre un 4 y un 5, pero esa diferencia puede estar menos clara en una persona sin estudios, con lo que se introduce, de este modo, error en la evaluación.
– Las instrucciones a la hora de rellenar el cuestionario. A la hora de rellenar un cuestionario las instrucciones que se den para hacerlo deben

ser claras y comprensibles. Estas deben adaptarse al grupo que se desee evaluar para que el criterio empleado a la hora de responder esté claro.

c) Factores relacionados con las personas. En este apartado entrarían todas las características personales de aquellos que van a responder al cuestionario (capacidad lectora, capacidad intelectual, capacidad discriminativa, estado emocional, etc.). Es necesario hacer especial mención a las situaciones en las que la persona está en un proceso de selección, ya que tratará de dar una imagen distorsionada de sí misma, tratando de adaptarse a lo que piensa que el seleccionador busca.

3.2. Procedimiento

Aunque en los estándares de la APA (1999) aparece este indicio de validez, apenas aportan información sobre cómo determinar si un test tiene indicios de validez basados en el proceso de respuesta. Dentro de las alternativas que proponen se encuentran técnicas como:

- *La entrevista*. En ella se les preguntará a las personas que responden al test por las diferentes estrategias empleadas para responder a cada uno de los ítems. El conocimiento de dichas estrategias puede conducir incluso al enriquecimiento de la definición del constructo estudiado.
- *Técnicas de pensamiento en voz alta*. Se les pide a las personas que rellenen el cuestionario diciendo en voz alta los diferentes procesos por los que pasan a medida que deben contestar el test.
- *Entrevistas cognitivas*. Están diseñadas para comprender cómo las personas que responden a un test comprenden la pregunta, recuperan la información relevante para responder, evalúan la relevancia de lo recordado y responden a la pregunta. Empleando esta información se pueden identificar potenciales errores de respuesta así como patrones de interpretación de las preguntas. También puede aportar información sobre los factores socioculturales que afectan al modo de responder (Czaja y Blair, 1996).

4. Evidencia de validez basada en la estructura interna

4.1. Concepto

Para la elaboración de un test se utilizarán distintos ítems o preguntas; es posible que se considere que todos los ítems son igual de relevantes para medir la característica estudiada, en cuyo caso obtendremos una puntuación total del test a partir de la simple suma de las puntuaciones obtenidas por el sujeto en los diferentes ítems.

La situación puede no ser tan sencilla cuando suponemos que no todos los ítems tienen la misma importancia en la medida del constructo, y por tanto será necesario ponderar las puntuaciones de los ítems antes de proceder a la suma; hablamos en este caso de puntuaciones compuestas. En esta situación la estructura del test que tendremos que determinar es unidimensional, ya que suponemos que todos los ítems, aunque de distinta manera, contribuyen a la medida de un único aspecto de la variable criterio.

Un test también puede presentar una estructura interna multidimensional, esto es, que las diferentes preguntas no miden un solo aspecto o dimensión sino dos o más.

La técnica estadística del análisis factorial nos servirá para el estudio de la contribución de los diferentes ítems a un solo factor (estructura unidimensional) o a varios factores (estructura multidimensional).

La técnica del análisis factorial nos permitirá determinar k factores subyacentes, a partir de una serie p de puntuaciones determinadas por los ítems iniciales del test. La idea es la búsqueda de un modelo parsimonioso (simple) a partir de un conjunto complejo de datos.

A partir de los trabajos de Spearman a principios del siglo XX, y sobre todo de Thurstone en los años cuarenta, el análisis factorial se evidencia como una buena herramienta en psicología para tratar de identificar los factores que intervienen en la inteligencia. Thurstone propuso la utilización del análisis factorial para dar explicación a las correlaciones que observaba entre diferentes ítems de los tests de inteligencia. Así, el empleo de esta técnica le permitió la identificación y diferenciación de las capacidades espacial, verbal y numérica, como factores de la inteligencia.

El problema de esta técnica estriba en las dificultades del cálculo, sobre todo a partir de un número elevado de ítems (variables); sin embargo, el desarrollo y la popularización actual de los programas estadísticos han permitido la difusión de esta y otras técnicas de análisis de datos multivariables.

4.2. Procedimientos

El término de *análisis factorial* no es un concepto unitario, sino que reúne diferentes procedimientos que persiguen la reducción inicial de múltiples variables en un menor número de factores. En procesos exploratorios, la más utilizada es la técnica del análisis en componentes principales, pero existen otras formas de extracción de los factores o componentes.

4.2.1. Unidimensionalidad

El análisis en componentes principales parte inicialmente de la matriz de correlaciones entre las diferentes variables. Disponemos de la matriz de correlaciones obtenida a partir de la administración de un test a una muestra de 52 individuos, y compuesto por ocho preguntas o ítems que intentan medir un único constructo, en este caso la autoestima de los sujetos.

Tabla 2. *Correlation matrix*

	ítem1	ítem2	ítem3	ítem4	ítem5	ítem6	ítem7	ítem8
ítem1	1,000	,447	,411	,444	,533	,337	,365	,442
ítem2	,447	1,000	,561	,662	,707	,528	,333	,522
ítem3	,411	,561	1,000	,665	,699	,462	,572	,540
ítem4	,444	,662	,665	1,000	,682	,518	,560	,564
ítem5	,533	,707	,699	,682	1,000	,467	,592	,488
ítem6	,337	,528	,462	,518	,467	1,000	,424	,418
ítem7	,365	,333	,572	,560	,592	,424	1,000	,422
ítem8	,442	,522	,540	,564	,488	,418	,422	1,000

La matriz de correlaciones presenta, en esta situación, una distribución de valores suficientemente uniforme, y no se detectan en este caso agrupaciones de variables con correlaciones altas entre ellas y bajas con las demás.

La varianza de cada variable (ítem) es posible descomponerla en tres fuentes de variación: la varianza factorial común, que comparten las variables en común, la varianza específica, o no compartida por otras variables, y la varianza del error. La varianza común, también denominada comunalidad (h^2), interesa que sea suficientemente alta una vez que hemos seleccionado los factores relevantes.

En el inicio del análisis la comunalidad de las variables es la unidad; después del análisis, cuanto más próxima esté a uno, más relación habrá con el factor o factores extraídos.

Tabla 3. Comunalidades

	Inicial	Extracción
ítem1	1,000	,410
ítem2	1,000	,628
ítem3	1,000	,670
ítem4	1,000	,722
ítem5	1,000	,743
ítem6	1,000	,455
ítem7	1,000	,489
ítem8	1,000	,518

Método de extracción: análisis de componentes principales

La comunalidad de un ítem *j* viene representada por:

$$h_j^2 = a_j^2 + b_j^2 + \cdots + k_j^2$$

Donde a_j^2, b_j^2,..., k_j^2 representan el cuadrado de los coeficientes de saturación de cada ítem con cada factor *A*, *B*,..., *K*, extraídos, siendo el coeficiente de saturación la correlación de cada ítem con los factores extraídos.

En el ejemplo, las variables que presentan mayor comunalidad son los ítems 4 y 5.

Tabla 4. Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4,636	57,947	57,947	4,636	57,947	57,947
2	,718	8,969	66,916			
3	,681	8,513	75,429			
4	,572	7,155	82,584			
5	,558	6,969	89,553			
6	,344	4,303	93,856			
7	,304	3,803	97,659			
8	,187	2,341	100,000			

Método de extracción: análisis de componentes principales

A partir de *p* variables, el análisis factorial extrae el mismo número de factores, no relacionados entre sí, y cada uno de los factores se define como combinación lineal de las *p* variables originales. Estos *p* factores se ordenan por orden de importancia; en efecto, el primer componente o factor es el mejor resumen de las relaciones lineales que presentan los datos. El segundo factor es la segunda mejor combinación de las variables, con la condición de que sea ortogonal (sin relación) con el primero, y así sucesivamente con el resto de los *p* factores o componentes.

Un criterio muy extendido, el más utilizado a la hora de extraer los componentes es el del valor propio o autovalor superior a 1. Otro sería retener los factores necesarios hasta conseguir un porcentaje adecuado de variabilidad explicada por los componentes.

El valor propio o autovalor (λ) se define como la suma de los cuadrados de las saturaciones o correlaciones de cada ítem con el componente en cuestión. Representa, por tanto, una medida de la variabilidad explicada en las variables por parte del componente o factor.

En la tabla de varianza total explicada, donde se desglosan los diferentes componentes, vemos que solo hay un componente con valor propio (4,63) superior a 1. Por tanto, confirmaría una estructura unidimensional del test, ya que todas las preguntas confluyen en un solo componente, identificable como el

constructo subyacente que se pretende medir. En nuestro ejemplo, las diferentes preguntas de la escala elaborada contribuirían a la medida de un único constructo psicológico que identificaríamos con la autoestima.

Tabla 5. Matriz de componentes[a]

	Componente
	1
ítem1	,640
ítem2	,793
ítem3	,819
ítem4	,850
ítem5	,862
ítem6	,675
ítem7	,699
ítem8	,720

Método de extracción: análisis de componentes principales.
[a]. 1 componentes extraídos

La matriz de componentes indica las correlaciones entre cada ítem con el componente, lo que hemos denominado anteriormente saturaciones (*factor loadings*). En el ejemplo vemos que los valores son altos, y además con poca fluctuación, lo cual indicaría que todos los ítems tienen similar importancia en la medida del constructo.

Con todos los indicadores mencionados podemos comprobar que la comunalidad de cada ítem, al haber seleccionado un solo factor, simplemente es el cuadrado de la saturación entre ítem y factor.

Así, para el ítem 1 la comunalidad final es $h_1^2 = (0{,}64)^2 = 0{,}41$. Un 41% de la variabilidad del primer ítem viene explicada por el componente.

El valor propio del componente 1 se obtiene de la suma de los cuadrados de las saturaciones de cada ítem con el factor.

Así, en el primer componente, $\lambda_1 = (0{,}64)^2 + (0{,}793)^2 + \ldots + (0{,}72)^2 = 4{,}63$.

Como tenemos 8 ítems, el máximo serían 8 componentes. Si hacemos el cociente 4,63 / 8 = 0,5795. Un 57,95% de la variabilidad total es explicada por el primer componente.

Una vez hemos extraído los componentes, dispondremos de la matriz de puntuaciones factoriales que nos proporciona las ponderaciones de cada variable para el cálculo de la puntuación de cada sujeto en los factores extraídos. Las puntuaciones factoriales (*factor scores*) para los datos individuales se calculan a partir de la matriz de coeficientes de puntuaciones factoriales

$$F_i = a_1 \cdot Z_1 + a_2 \cdot Z_2 + \cdots + a_p \cdot Z_p$$

a_i: Coeficientes de ponderación de cada variable para cada factor.

Z_i: Puntuaciones tipificadas de los valores de cada variable, obtenidos por cada individuo.

Con los datos del ejemplo la matriz de ponderaciones:

Tabla 6. Matriz de coeficientes para el cálculo de las puntuaciones en las componentes

	Componente
	1
ítem1	,138
ítem2	,171
ítem3	,177
ítem4	,183
ítem5	,186
ítem6	,146
ítem7	,151
ítem8	,155

Método de extracción: análisis de componentes principales

Para cada sujeto es posible calcular una puntuación factorial del índice de autoestima:

$$F_1 = 0{,}138 \cdot Z_{ítem1} + 0{,}171 \cdot Z_{ítem2} + \cdots + 0{,}155 \cdot Z_{ítem8}$$

De todos modos, en este ejemplo se observa que las ponderaciones de todos los ítems son muy similares, por lo que la contribución de todas las preguntas es muy similar en la medida del constructo de interés; por tanto, sería adecuado optar por una puntuación simple, únicamente sumando las puntuaciones obtenidas en cada ítem, frente a una puntuación compuesta ponderando los valores de cada ítem.

4.2.2. Multidimensionalidad

En muchas ocasiones, aunque de entrada intentemos elaborar una escala para la medida de un solo constructo psicológico, es posible que después de la primera administración del test, en la prueba piloto, observemos que en realidad los ítems se agrupan entre ellos y afectan a diferentes constructos subyacentes.

Presentamos otro ejemplo en el que se ha realizado un cuestionario con la intención de medir las actitudes sobre ideas religiosas en una muestra de 870 sujetos. Los ítems se han identificado con el concepto principal que implicaba la pregunta. La matriz de correlaciones de Pearson entre los ítems se presenta a continuación:

Tabla 7. *Correlation matrix*

	Sent. vida	Religión	Obediencia	Más allá	Experi.	Inseguridad	Influencia	Independ.
Sentido de la vida	1,000	,295	,220	,253	,226	,134	,178	,027
Religión	,295	1,000	,440	,507	,243	,099	,241	,103
Obediencia	,220	,440	1,000	,339	,292	,063	,336	,054
Más allá	,253	,507	,339	1,000	,309	,113	,278	,121
Experiencia	,226	,243	,292	,309	1,000	,078	,204	,049
Inseguridad	,134	,099	,063	,113	,078	1,000	,117	,169
Influencia	,178	,241	,336	,278	,204	,117	1,000	,102
Independencia	,027	,103	,054	,121	,049	,169	,102	1,000

Las correlaciones fluctúan en un rango similar entre las diferentes preguntas; quizá las que presentan correlaciones inferiores son las preguntas referidas a inseguridad e independencia.

Al aplicar el correspondiente análisis en componentes principales, y utilizando el criterio de valor propio superior a 1, para la extracción de los componentes, obtenemos el siguiente listado:

Tabla 8. Comunalidades

	Inicial	Extracción
Sentido de la vida	1,000	,279
Religión	1,000	,554
Obediencia	1,000	,513
Más allá	1,000	,525
Experiencia	1,000	,333
Inseguridad	1,000	,562
Influencia	1,000	,315
Independencia	1,000	,588

Método de extracción: análisis de componentes principales

Tabla 9. Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,554	31,926	31,926	2,554	31,926	31,926
2	1,115	13,936	45,862	1,115	13,936	45,862
3	,901	11,258	57,120			
4	,826	10,329	67,448			
5	,787	9,840	77,288			
6	,739	9,234	86,522			
7	,632	7,906	94,428			
8	,446	5,572	100,000			

Método de extracción: análisis de componentes principales

Tabla 10. Matriz de componentes[a]

	Componente	
	1	2
Sentido de la vida	,526	-,046
Religión	,735	-,115
Obediencia	,685	-,208
Más allá	,723	-,056
Experiencia	,558	-,146
Inseguridad	,267	,701
Influencia	,560	,044
Independencia	,221	,734

Método de extracción: análisis de componentes principales.
[a]. 2 componentes extraídos

De los ocho componentes posibles, solamente los dos primeros cumplen el criterio de autovalor superior a 1, aunque un tercer factor está a punto de llegar a este límite ($\lambda_3 = 0{,}901$). En todo caso, se extraen dos componentes.

La última de las tablas presentadas (matriz de componentes) nos muestra las saturaciones o correlaciones entre los diferentes ítems y los dos componentes.

Recordemos que la comunalidad de los ítems representa la variabilidad explicada del ítem por los factores extraídos. Así, en el ítem 1 (sentido de la vida):

$$h_1^2 = (0{,}526)^2 + (-0{,}046)^2 = 0{,}279$$

Explicaría un 27,9% de la variabilidad del ítem. Es la más baja de todos los ítems del test, quizá este ítem saturaría con un tercer componente.

El valor propio (autovalor) de cada componente se calcula con la suma de cuadrados de las correlaciones (saturaciones) entre ítems y componente.

Componente 1: $\lambda_1 = (0{,}526)^2 + (0{,}735)^2 + \ldots + (0{,}221)^2 = 2{,}554$

Componente 2: $\lambda_2 = (-0{,}046)^2 + (-0{,}115)^2 + \ldots + (0{,}734)^2 = 1{,}115$

El primer componente (2,554 / 8 = 0,319) explicaría un 31,9% de la variabilidad presentada por los ítems, mientras que el segundo (1,115 / 8 = 0,139) explicaría un 13,9%. Un total de un 45,8% de la varianza total es explicada por la combinación de los dos factores o componentes.

El análisis de la matriz de saturaciones nos permitirá intentar buscar interpretación a los dos componentes, en función de los ítems que correlacionen de forma más alta.

En el ejemplo observamos que el primer componente tiene altas saturaciones en los ítems 1, 2, 3, 4, 5, y 7; mientras que las correlaciones son bajas en los ítems 6 y 8. En el segundo componente ocurre justo al contrario: tiene altas correlaciones con los ítems 6 y 8, y en cambio bajas en los demás.

En ocasiones la solución final no presenta, a simple vista, una tan fácil interpretación; en estos casos es posible utilizar una rotación de los ejes para conseguir que las correlaciones sean fuertes en un eje o componente y bajas en los demás. Recordemos que los ejes son ortogonales y no relacionados entre ellos. Una de las técnicas matemáticas de rotación de los ejes es la rotación varimax, la más utilizada en procesos exploratorios, aunque los programas estadísticos incorporan otras, como las rotaciones quartimax, equimax, etc.

La matriz de saturaciones con la solución rotada con el método varimax, con los datos del ejemplo, quedaría como sigue:

Tabla 11. Matriz de componentes rotados[a]

	Componente	
	1	2
Sentido de la vida	,522	,080
Religión	,742	,062
Obediencia	,715	-,040
Más allá	,715	,117
Experiencia	,577	-,010
Inseguridad	,093	,744
Influencia	,533	,175
Independencia	,041	,765

Método de extracción: análisis de componentes principales. Método de rotación: normalización Varimax con Kaiser.
[a]. La rotación ha convergido en 3 iteraciones.

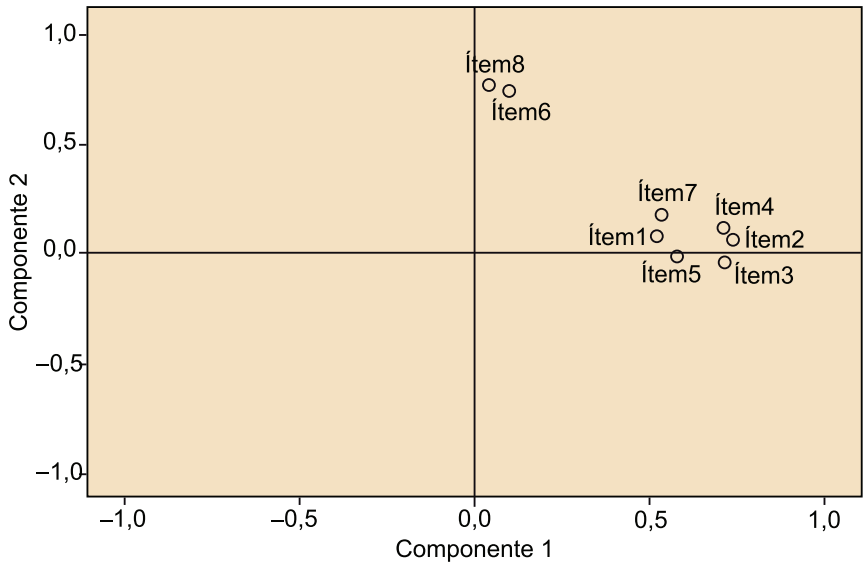
Vemos cómo la solución rotada confirma en este caso la conclusión previa, seis de las preguntas saturan en el primer componente. Viendo los seis ítems

que saturan este componente –sentido de la vida, religión, obediencia, más allá, experiencia, influencia–, podemos interpretar este componente como la medida de la actitud sobre ideas religiosas, que era el motivo inicial de la elaboración del test.

El segundo componente, se supone que no esperado inicialmente en la elaboración del cuestionario, satura los ítems inseguridad e independencia. Este segundo componente se puede interpretar como una medida de la actitud hacia la dependencia personal.

Al tratarse de solo dos componentes es posible representar gráficamente en un espacio bidimensional los dos ejes que representan los componentes y la situación de los ítems respecto a ellos, en función de las saturaciones obtenidas.

Figura 5. Gráfico de componentes en espacio rotado



El gráfico da una rápida información visual de la agrupación de los ítems en los dos componentes.

Una vez obtenidos los factores terminales, podremos calcular los valores o las puntuaciones factoriales de las dimensiones teóricas que hemos asociado a los respectivos factores, ideas religiosas (componente 1) y dependencia (componente 2).

Tabla 12. Matriz de coeficientes para el cálculo de las puntuaciones en las componentes

	Componente	
	1	2
Sentido de la vida	,210	,009
Religión	,304	-,032
Obediencia	,305	-,118
Más allá	,287	,018
Experiencia	,243	-,076
Inseguridad	-,047	,635
Influencia	,204	,090
Independencia	-,072	,660

Método de extracción: análisis de componentes principales. Método de rotación: normalización Varimax con Kaiser.

Para cada sujeto *i* es posible calcular las puntuaciones factoriales de las nuevas variables ideas religiosas y dependencia:

$$Ideas\ religiosas_i = 0,21 \cdot Zi_{ítem1} + 0,304 \cdot Zi_{ítem2} + \cdots + 0,204 \cdot Zi_{ítem7} - 0,072 \cdot Zi_{ítem8}$$

$$Dependencia_i = 0,009 \cdot Zi_{ítem1} - 0,032 \cdot Zi_{ítem2} + \cdots + 0.,09 \cdot Zi_{ítem7} + 0,66 \cdot Zi_{ítem8}$$

Los diferentes paquetes estadísticos permiten el cálculo automático y la generación de estas nuevas variables generadas, y que representarían la medida de los constructos subyacentes.

5. Evidencia de validez basada en la relación con otras variables

5.1. Concepto

En el proceso de validación de una nueva prueba psicológica podemos ayudarnos de la existencia de otros instrumentos de medida del constructo de interés, que estén contrastados como válidos y fiables. En este proceso hablaremos de validez convergente o correlación entre puntuaciones del test con otras medidas del mismo constructo realizadas a partir de diferentes técnicas o indicadores.

Las diferentes técnicas estadísticas de relación entre variables nos servirán para determinar el coeficiente de validación entre las dos variables. Así, el más utilizado será el coeficiente de correlación de Pearson, en el caso de que las dos variables sean cuantitativas, pero también cualquiera de sus variaciones, como los coeficientes de Spearman, Biserial puntual, Biserial, phi, Tetracórica, etc., en función de cómo sean las dos variables que hay que relacionar.

Tabla 13

Variable A	Variable B	Coeficiente correlación
Numérica (intervalo o razón)	Numérica (intervalo o razón)	*r* de Pearson
Numérica (intervalo o razón)	Numérica (ordinal)	r_s Spearman o T de Kendall
Numérica (ordinal)	Numérica (ordinal)	r_s Spearman o T de Kendall
Cualitativa	Cualitativa	V de Cramer
Cualitativa (dicotómica)	Numérica (intervalo o razón)	*rb* Biserial o r_{bp} Biserial puntual
Cualitativa (dicotómica)	Cualitativa (dicotómica)	ϕ Phi o r_t Tretacórica

Para poder conocer el nivel de bienestar físico y psicológico en personas mayores, nos interesa validar un nuevo cuestionario que hemos elaborado para determinar el grado de independencia en las actividades básicas de la vida diaria

(ABVD). A tal objeto, podemos utilizar algunas de las pruebas que ya existen en el mercado y que se encuentran suficientemente contrastadas. En una muestra de 300 sujetos mayores de 70 años, e ingresados en centros geriátricos, administramos la nueva prueba elaborada junto con la escala de medida de independencia funcional (FIM) (Keith, Granger, Hamilton y Sherwin, 1987) y la escala de grado de autonomía de Barthel (Mahoney y Barthel, 1965).

En la tabla siguiente se muestra la matriz de correlaciones de Pearson (datos simulados) entre las tres pruebas administradas:

Tabla 14

	ABVD	FIM	Barthel
ABVD	1		
FIM	0,69	1	
Barthel	0,67	0,77	1

Los valores de la correlación entre la nueva prueba (ABVD) y las escalas FIM y Barthel presentan valores suficientemente altos (0,69 y 0,67, respectivamente), lo cual indica una alta validez concurrente de la nueva prueba elaborada con las técnicas previas para la medida del grado de autonomía de las personas mayores analizadas.

5.2. Evidencia de decisión (sensibilidad y especificidad)

En situaciones en las que la prueba realizada tenga como objetivo el diagnóstico o la clasificación de los sujetos en dos grupos (diagnóstico negativo-diagnóstico positivo) hablaremos de la validez de decisión cuando comparemos esta nueva prueba con otro método de diagnóstico anterior suficientemente contrastado. En la validez de decisión podemos distinguir dos procesos: por una parte, la sensibilidad de la prueba, definida como la capacidad de esta en la detección de verdaderos positivos, y por otra parte, la especificidad, definida como la capacidad de determinación de diagnósticos negativos verdaderos.

Tabla 15

		Diagnóstico prueba inicial		
		Positivo	**Negativo**	**Total**
Diagnóstico nueva prueba	**Positivo**	Decisión correcta (f_{11})	Falso positivo (f_{12})	$f_{1.}$
	Negativo	Falso negativo (f_{21})	Decisión correcta (f_{22})	$f_{2.}$
	Total	$f._1$	$f._2$	n

Una medida del acuerdo logrado a través de las dos pruebas diagnósticas consistirá en calcular el porcentaje de acuerdo (P_c) entre ambas técnicas a partir de la razón entre la suma de decisiones correctas y el total de decisiones.

$$P_c = \frac{f_{11} + f_{22}}{n}$$

La sensibilidad de la nueva prueba la obtendremos a partir de la proporción de sujetos clasificados correctamente como verdaderos positivos.

$$Sensibilidad = \frac{diagnósticos\ positivos\ de\ la\ prueba}{total\ diagnósticos\ positivos} = \frac{f_{11}}{f_{.1}}$$

Mientras que la especificidad se obtiene mediante el cociente de los diagnosticados sin trastorno por la prueba entre el total de diagnósticos negativos.

$$Especificidad = \frac{diagnósticos\ negativos\ de\ la\ prueba}{total\ diagnósticos\ negativos} = \frac{f_{22}}{f_{.2}}$$

Un índice global para valorar la validez lo proporciona el cálculo del coeficiente kappa, establecido inicialmente como indicador del acuerdo entre dos observadores. La ventaja que presenta consiste en su fácil interpretación, similar a la de otros indicadores de relación entre variables. En efecto, su valor fluctúa entre 0 (ningún acuerdo) a valor 1 (máximo acuerdo). Su fórmula de cálculo es sencilla:

$$K = \frac{F_c - F_a}{n - F_a}$$

Donde

$$F_c = f_{11} + f_{22} \quad y \quad F_a = \frac{f_{1.} \cdot f_{.1} + f_{2.} \cdot f_{.2}}{n}$$

Tabla 16. Criterios Alman para interpretar kappa

Valor	Relación
0-0,20	Inexistente
0,21-0,40	Muy baja
0,41-0,60	Moderada
0,61-0,80	Buena
0,81-1	Intensa

En una consulta psicológica se pretende validar una nueva prueba, más simple que las tradicionales, para el diagnóstico de trastorno de depresión de los pacientes atendidos. En una muestra de 500 pacientes atendidos en el centro se administran dos pruebas (tradicional y versión breve) para el diagnóstico del trastorno de depresión.

Tabla 17

		Diagnóstico tradicional depresión		
		Positivo	Negativo	Total
Versión breve Escala Hamilton	Positivo	125	50	175
	Negativo	25	300	325
	Total	150	350	500

$$P_c = \frac{125 + 300}{500} = 0,85$$

Las dos pruebas presentan un porcentaje de acuerdo (85%) elevado.

Asimismo, los valores de sensibilidad y especificidad indican buena capacidad de la nueva prueba en la detección de sujetos con trastorno depresivo (sensibilidad = 0,83), como en la detección de los sujetos sin depresión (especificidad = 0,86).

$$Sensibilidad = \frac{125}{150} = 0,83$$

$$Especificidad = \frac{300}{350} = 0,86$$

$$F_c = 125 + 300 = 425 \quad y \quad F_a = \frac{175 \cdot 150 + 325 \cdot 350}{500} = 280$$

$$K = \frac{425 - 280}{500 - 280} = 0,66$$

El cálculo del índice kappa de acuerdo entre las dos pruebas (K = 0,66) indica una buena relación entre las dos pruebas. Por tanto, parece adecuado que la economía de tiempo y esfuerzo (tanto para el paciente como para el terapeuta) justificaría la nueva escala de diagnóstico de depresión en función de los resultados obtenidos en la validez de decisión.

5.3. Evidencias convergentes y/o discriminantes

Hasta ahora, en este apartado, nos hemos referido a escalas que pretenden mesurar un solo constructo psicológico. Si nos referimos a pruebas formadas por ítems que miden diferentes constructos (múltiples rasgos), podremos diferenciar entre dos tipos de validez. Por una parte, la validez convergente (enunciada anteriormente), es decir, la validez que determinan diferentes pruebas que miden el mismo constructo, y por otra parte, la validez discriminante, que viene determinada por la medida de diferentes constructos dentro de la misma prueba.

La matriz de correlaciones entre las puntuaciones de los diferentes rasgos obtenidos a partir de las diferentes escalas (matriz multirrasgo-multimétodo; Campbell y Fiske, 1959) nos servirá para determinar los diferentes valores de validez convergente y discriminante.

Imaginemos la situación en la que disponemos de tres escalas diferentes, formadas por ítems que miden los mismos dos constructos subyacentes. Escalas A, B y C, que miden los constructos 1 y 2. Para cada sujeto analizado tendremos por tanto seis puntuaciones diferentes obtenidas por la combinación de cada prueba y cada rasgo analizado.

Tabla 18

	A1	A2	B1	B2	C1	C2
A1	**Fi**					
A2	Vd	**Fi**				
B1	Vc		**Fi**			
B2		Vc	Vd	**Fi**		
C1	Vc		Vc		**Fi**	
C2		Vc		Vc	Vd	**Fi**

En la tabla anterior las diferentes escalas se representan por las letras A, B y C; dentro de cada escala 1 y 2 representan los dos constructos que analizar.

Si observamos la matriz multirrasgo-multimétodo (MRMM), encontramos en la diagonal principal valores de fiabilidad de las escalas (valores 1 si son obtenidas en una única administración).

Los valores de validez convergente se encuentran en las combinaciones de los mismos rasgos y diferentes escalas (por ejemplo, casilla A1 y B1), esperando que estos valores de validez sean suficientemente altos, lo cual indica convergencia de las diferentes maneras de medir el constructo, aportando evidencia real de la existencia del constructo.

Los valores de validez discriminante serán los coeficientes de correlación obtenidos dentro de la misma escala por las puntuaciones de los diferentes rasgos.

En este caso, esperamos que los diferentes constructos sean suficientemente independientes entre sí para que las correlaciones sean próximas a cero.

En una muestra de 600 sujetos se han utilizado tres pruebas diferentes de personalidad (tests 1, 2 y 3), formadas cada una de ellas por ítems referidos a los tres mismos constructos de la personalidad (rasgos A, B y C).

La tabla siguiente muestra los valores de la matriz de correlaciones entre las 9 variables obtenidas de la combinación de las tres pruebas y los tres rasgos (3 x 3).

Taula 19

	A1	**B1**	**C1**	**A2**	**B2**	**C2**	**A3**	**B3**	**C3**
A1	1								
B1	*0,03*	1							
C1	*0,28*	*0,17*	1						
A2	0,73	0,14	0,22	1					
B2	0,15	0,69	0,03	*0,18*	1				
C2	0,12	0,04	0,81	*0,03*	*0,15*	1			
A3	0,77	0,11	0,09	0,77	0,21	0,16	1		
B3	0,21	0,75	0,18	0,21	0,68	0,03	*0,15*	1	
C3	0,19	0,05	0,78	0,22	0,09	0,72	*0,14*	*0,09*	1

Los valores de validez convergente son los valores resaltados por el subrayado. Los valores de validez discriminante se resaltan por la cursiva.

Si nos fijamos en el rasgo B, se detecta convergencia a partir de la medida a través de diferentes pruebas:

$$r\,(B1 - B2) = 0{,}69$$

$$r\,(B1 - B3) = 0{,}75$$

$$r\,(B2 - B3) = 0{,}68$$

Si nos fijamos en la escala 1, observamos que existe suficiente independencia entre las diferentes medidas de los tres rasgos medidos:

$$r\,(A1 - B1) = 0{,}03$$

$$r\,(A1 - C1) = 0{,}28$$

$$r\,(B1 - C1) = 0{,}17$$

5.4. Evidencias basadas en las relaciones test-criterio

En ocasiones, un test o prueba psicológica construida para la medida de determinado constructo psicológico puede encontrarse relacionada con otra variable de interés, que se denomina criterio.

Por ejemplo, imaginemos que hemos elaborado una prueba válida que nos permite medir la capacidad de razonamiento numérico de las personas (test), y observamos que presenta una muy alta correlación con los resultados que obtienen los sujetos en una determinada prueba de matemáticas (criterio).

Podemos distinguir tres tipos de situaciones:

- Validez concurrente o simultánea.
- Validez predictiva.
- Validez retrospectiva.

5.4.1. Validez concurrente o simultánea

En este caso el test y el criterio se miden de manera simultánea. Obtendremos validez concurrente al obtener valores altos de coeficientes de correlación entre las puntuaciones del test y del criterio. Por tanto, nos permite validar el test, inicialmente elaborado para la medida de otra variable, para la medida del criterio.

En función del tipo de escala de medida utilizado tanto para las puntuaciones del test como del criterio, utilizaremos un tipo u otro de coeficiente para la medida de la correlación.

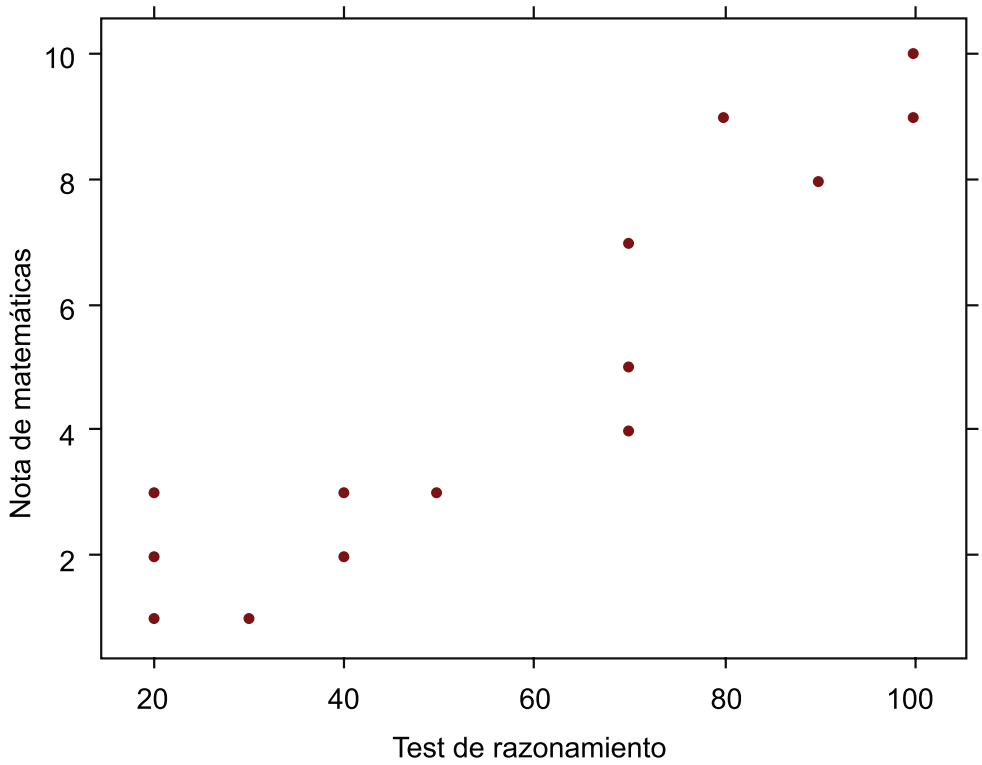
Como ejemplo veamos la siguiente situación, en la que a un grupo de 20 sujetos se les ha administrado un test de razonamiento numérico, justo antes de realizar determinada prueba de matemáticas.

Tabla 20

Test razonamiento	Nota matemáticas
100	10
100	9
100	9
90	8
90	8
80	9
70	7
70	7
70	7
70	5
70	4
70	4
50	3
40	3
40	2
40	2
30	1
20	1
20	3
20	2

Una visión del gráfico de dispersión nos dará una idea de si se observa relación lineal entre ambas pruebas o no:

Figura 6. Gráfico de dispersión (nube de puntos)



Al calcular el coeficiente de correlación de Pearson:

```
> rcorr.adjust(Datos[,c("Nota.Matemáticas",
"Test.Razonamiento")], type="pearson")
                Nota.Matemáticas     Test.Razonamiento
Nota.Matemáticas        1.00          0.92291
Test.Razonamiento       0.92291       1.00
n= 20
```

Se obtiene un valor de validez concurrente igual a 0,92291, que indica una fuerte relación directa y próxima a 1.

Una medida de la bondad de ajuste lineal entre las dos variables se define por r^2, es decir, el valor del cuadrado de la correlación, en nuestro ejemplo $r^2 = (0{,}92291)^2 = 0{,}8518$. Este valor, que se denomina coeficiente de determinación, multiplicado por 100, indica el porcentaje de variabilidad en la variable criterio, que viene explicado por la relación con la variable independiente. Por tanto, el 85,18% de la variabilidad que presentan las puntuaciones obtenidas en la nota de matemáticas estaría explicada por la relación que presenta con los valores obtenidos en el test de razonamiento numérico.

5.4.2. Validez predictiva

Si conocemos que un determinado test y una variable criterio se encuentran altamente relacionados, será posible utilizar los valores obtenidos en el test para la predicción o el pronóstico de los valores que se obtendrán en el criterio. Hablaremos en este caso de la validez predictiva que tiene el test respecto a la variable criterio.

Por ejemplo, si queremos seleccionar un candidato para un determinado puesto de trabajo, podemos utilizar determinadas pruebas que tengan alta validez predictiva con el futuro rendimiento de los candidatos en el puesto de trabajo. O, utilizando el ejemplo mencionado anteriormente, podemos hacer un pronóstico de la nota que obtendrán los sujetos en una determinada prueba de matemáticas, a partir de las puntuaciones que obtuvieron en su momento, en el test de razonamiento numérico.

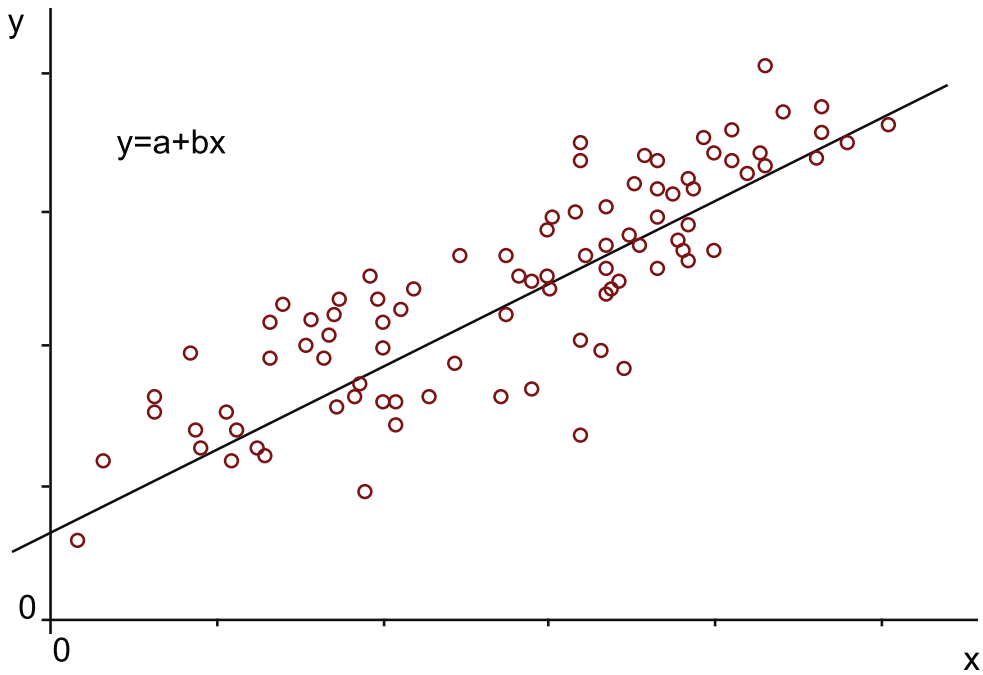
Cuando tanto las puntuaciones del test como el criterio son puntuaciones numéricas y hemos calculado el correspondiente coeficiente de correlación de Pearson, siendo este estadísticamente significativo, es posible establecer un modelo de regresión lineal para poder realizar el pronóstico de los valores del criterio.

Las puntuaciones en el test constituyen la variable independiente del modelo (variable predictora), mientras que el criterio representa la variable dependiente.

Regresión lineal simple

Es el caso más sencillo: solo disponemos de una variable independiente (X) y una variable dependiente (Y).

Figura 7. Gráfico de dispersión con recta de regresión



La regresión lineal describe una relación lineal entre Y e X, esto es, representar una recta en el gráfico de dispersión, que mejor ajuste a la nube de puntos.

La expresión de una línea recta es $y = a + bx$, donde b representa la pendiente de la recta, o sea, el cambio que se produce en y a partir del cambio que se produzca en x; y a se denomina intersección o intercepta, y es el valor que toma y cuando x es igual a cero.

Para encontrar los coeficientes de la regresión, a y b, usamos un método de estimación muy conocido en estadística, el método de mínimos cuadrados, que minimiza la suma de los cuadrados de las diferencias (o residuos) entre los valores y_i y los valores estimados según la recta de regresión $y'_i = a + bx_i$.

A partir de los datos (x_i, y_i), $i = 1, ..., n$, estimamos los coeficientes a y b de la recta de regresión. Así pues, tenemos:

– Pendiente:

$$b = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{s_x^2}$$

S_{xy}: Covarianza entre x e y.

$s^2{}_x$: Varianza de x.

– Intersección:

$$a = \overline{y} - b \cdot \overline{x}$$

Comparando las fórmulas de la pendiente b y del coeficiente de correlación r, tenemos la relación siguiente:

$$b = r_{xy} \cdot \frac{s_y}{s_x}$$

Es posible verificar o validar el modelo de regresión a partir del coeficiente de determinación r^2. Recordemos que lo hemos definido anteriormente como una medida de bondad de ajuste, o medida de la proximidad de los puntos a la recta estimada. Representa la proporción de varianza de la variable dependiente explicada por la recta de regresión.

El valor $1 - r^2$ cuantifica la proporción de varianza que no es explicada por la regresión. A partir de estos dos valores podemos calcular un estadístico de contraste:

$$F_{EC} = \frac{r^2}{\left(1 - r^2\right) \Big/ n - 2}$$

Este estadístico de contraste F se distribuye siguiendo una distribución F de Snedecor, con 1 grado de libertad en el numerador y $n - 2$ grados de libertad en el denominador.

Las hipótesis que contrastar serán:

H_0: El modelo no es válido, no hay relación.

H_1: Sí existe relación, por tanto el modelo sí es válido.

Siguiendo con el ejemplo del test de razonamiento numérico y el criterio de la nota de matemáticas, el resultado con el programa R, es el siguiente:

```
Call:
lm(formula = Nota.Matemáticas ~ Test.Razonamiento, data = Datos)
Residuals:
     Min       1Q         Median       3Q      Max
-2.01102    -0.96970       -0.03857       0.98898  2.05785
Coefficients:
                     Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)         -1.085399   0.673899    -1.611     0.125
Test.Razonamiento    0.101377   0.009969    10.170     6.89e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.201 on 18 degrees of freedom
Multiple R-squared: 0.8518, Adjusted R-squared: 0.8435
F-statistic: 103.4 on 1 and 18 DF,  p-value: 6.89e-09
```

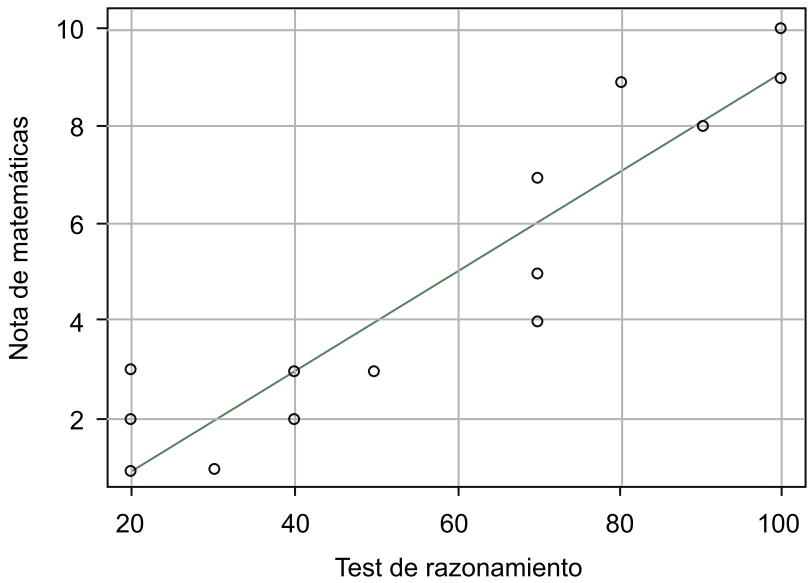
Consultado el listado, podemos especificar el modelo de regresión:

$$\textit{Nota matemáticas} = -1{,}085 + 0{,}1014 \times \textit{Test} + \textit{Residual}$$

También observamos que el modelo se encuentra verificado, ya que el valor de *p* (grado de significación) que acompaña al valor del estadístico de contraste (F = 103,4) es tendente a cero. Por tanto, nada se opone a rechazar la hipótesis nula, y el modelo se encuentra validado.

Si representamos en el gráfico de dispersión la recta de regresión:

Figura 8. Gráfico de dispersión con ajuste de la recta



A partir de la expresión de la recta de regresión, podemos realizar el pronóstico para cada sujeto del valor de nota de matemáticas en función del test, así como el cálculo del residual, calculado mediante la diferencia entre la puntuación real obtenida y la puntuación pronosticada.

Tabla 21

Test razonamiento	Nota matemáticas	Pronóstico	Residual
100	10	9,0546	0,9454
100	9	9,0546	-0,0546
100	9	9,0546	-0,0546
90	8	8,0406	-0,0406
90	8	8,0406	-0,0406

Test razonamiento	Nota matemáticas	Pronóstico	Residual
80	9	7,0266	1,9734
70	7	6,0126	0,9874
70	7	6,0126	0,9874
70	7	6,0126	0,9874
70	5	6,0126	-1,0126
70	4	6,0126	-2,0126
70	4	6,0126	-2,0126
50	3	3,9846	-0,9846
40	3	2,9706	0,0294
40	2	2,9706	-0,9706
40	2	2,9706	-0,9706
30	1	1,9566	-0,9566
20	1	0,9426	0,0574
20	3	0,9426	2,0574
20	2	0,9426	1,0574

```
> numSummary(Datos[,c("Nota.Matemáticas", "Pronóstico", "Residual")],
+ statistics=c("mean", "sd",”var”))
                     mean       sd        var        n
Nota.Matemáticas    5.2000    3.036619   9,221      20
Pronóstico          5.2014    2.803138   7,855      20
Residual           -0.0014    1.169176   1,367      20
```

Tal como hemos indicado anteriormente, se define el coeficiente de determinación como el cociente entre la varianza explicada por la regresión y la varianza total de la variable criterio:

$$r^2 = \frac{s^2_{y'}}{s^2_y} = 1 - \frac{s^2_{y-y'}}{s^2_y}$$

s^2_y : Varianza de la variable dependiente *Y*.

$s^2_{y'}$: Varianza de los pronósticos obtenidos mediante la ecuación de regresión.

$s^2_{y-y'}$: Varianza de los errores producidos.

Con los datos de nuestro ejemplo:

$$r^2 = \frac{7,855}{9,221} = 1 - \frac{1,367}{9,221} = 0,8518$$

Por tanto, a partir de los valores del coeficiente de determinación y de la varianza de la variable criterio, es posible, despejando de la expresión, obtener el valor de la varianza de los errores:

$$s^2_{y-y'} = s^2_y \left(1 - r^2\right)$$

La desviación típica de los errores o error típico o estándar del error nos ayudará en la estimación por intervalo de nuevos valores desconocidos.

$$s_{y-y'} = s_y \sqrt{1 - r^2}$$

En efecto, si necesitamos realizar un pronóstico en el criterio a partir de un nuevo valor en el test, o variable predictora, es posible realizarlo de forma puntual, pero conseguiremos mejores estimaciones, dada una probabilidad, si se realiza por intervalo.

$$IC^{1-\alpha} \rightarrow y' \pm t_{n-1;\alpha/2} \cdot s_{y-y'}$$

$1 - \alpha$: Nivel de confianza del intervalo construido.

t: Valor de la distribución t de Student-Fisher tabulado en función de α y de los grados de libertad (*n*-1).

Imaginemos que un nuevo sujeto obtiene una puntuación igual a 60 en el test de razonamiento numérico. La estimación puntual del valor en la nota de matemáticas será:

$$Nota\,Matemáticas' = -1,085 + 0,1014 \cdot 60 = 4,999$$

Si realizamos una estimación por intervalo, con un nivel de confianza del 95%:

$$IC^{0,95} \rightarrow 4,999 \pm 2,093 \cdot 1,169 = \left[2,552 \quad 7,446\right]$$

Con una probabilidad de 0,95 el valor en el criterio de un sujeto que obtenga una puntuación 60 en el test estará entre 2,552 y 7,446 puntos.

Regresión lineal múltiple

El modelo lineal general plantea que una variable dependiente (criterio) sea función de varias variables independientes, situación por otra parte bastante más habitual. En el caso de dos variables independientes la expresión que relaciona las tres variables será la fórmula de un plano. En las situaciones en las que existan más de dos variables independientes, situaciones multivariantes, hablaremos del hiperplano de regresión.

$$Criterio = a + b_1 \cdot Test_1 + b_2 \cdot Test_2 + b_3 \cdot Test_3 + \cdots + b_p \cdot Test_p + Residual$$

Recuperando el ejemplo anterior, imaginemos que a los 20 sujetos se les ha administrado un test de razonamiento junto con un test de cálculo mental, previamente a la realización de una prueba de matemáticas, que representa el criterio que más adelante queremos pronosticar.

Tabla 22

Test razonamiento	Test cálculo	Nota matemáticas
100	9	10
100	8	9
100	8	9
90	8	8
90	7	8
80	9	9
70	6	7
70	5	7
70	6	7
70	6	5
70	4	4
70	4	4
50	5	3
40	4	3
40	4	2
40	5	2
30	3	1
20	2	1
20	3	3
20	3	2

Al calcular la matriz de correlaciones de Pearson:

Tabla 23

	Test razonamiento	Test cálculo	Nota matemáticas
Test razonamiento	1		
Test cálculo	0,891763351	1	
Nota matemáticas	0,922905961	0,925261006	1

Observamos que el test de cálculo mental también se encuentra altamente correlacionado con la variable criterio (nota de matemáticas). Asimismo, vemos una alta correlación entre las dos variables independientes, test de razonamiento numérico y test de cálculo mental ($r = 0{,}89176$).

El análisis de regresión múltiple nos ayudará a determinar si la incorporación de esta nueva variable aumenta, significativamente, la variabilidad explicada por la regresión en la variable criterio.

El análisis de regresión se basará en el análisis de la relación conjunta entre la variable criterio y el conjunto de las dos variables independientes. El cuadrado de esta correlación múltiple será el nuevo coeficiente de determinación.

La verificación del modelo se realizará a partir de la expresión:

$$F_{EC} = \frac{r^2 / p}{\left(1 - r^2\right) / n - p -}$$

p es igual al número de variables independientes en el modelo.

El estadístico de contraste F se distribuye siguiendo una distribución F de Snedecor, con p grados de libertad en el numerador y $(n - p - 1)$ grados de libertad en el denominador.

A continuación se presenta el listado obtenido en este ejemplo mediante el uso del programa R:

```
> RegModel.2 < lm(Nota.Matemáticas~Test.Numérico +
Test.Razonamiento, data=Datos)

> summary(RegModel.2)

Call:

lm(formula = Nota.Matemáticas ~ Test.Numérico + Test.Razonamiento, data
= Datos)

Residuals:

       Min           1Q       Median       3Q         Max

-1.72686            -0.64006  -0.00108     0.52167    1.74005

Coefficients:

                  Estimate    Std.  Error    t value  Pr(>|t|)

Intercept)          -1.91579     0.62646    -3.058    0.00711 **

Test.Cálculo         0.70883     0.23719     2.988    0.00825 **

Test.Razonamiento    0.05246     0.01835     2.858    0.01088 *

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001  on 17 degrees of freedom

Multiple R-squared: 0.9028, Adjusted R-squared: 0.8914

F-statistic: 78.96 on 2 and 17 DF,  p-value: 2.481e-09
```

El valor de coeficiente de determinación es 0,9028; por tanto, un 90,28% de la varianza de la variable criterio viene explicada por la regresión entre esta variable y la combinación de las dos variables independientes.

La ecuación del plano de regresión se encuentra verificada globalmente, ya que el valor del estadístico de contraste ($F = 78{,}96$) indicia una probabilidad tendente a cero (*p value*) de que no exista relación entre las variables.

Es importante también observar si se encuentran significados los diferentes coeficientes de la regresión (b). En este caso, tanto el coeficiente que afecta al

test de razonamiento (*p value* = 0,01) como el que afecta al test de cálculo mental (p = 0,008), se encuentran verificados, ya que sus respectivos grados de significación asociados son próximos a cero.

La especificación de la ecuación resultante del modelo de regresión quedará, por tanto, de la siguiente manera:

$$\textit{Nota matemáticas} = -1{,}916 + 0{,}709 \cdot \textit{Test\_cálculo} + 0{,}052 \cdot \textit{Test\_razonamiento} + \textit{Residual}$$

El listado, asimismo, informa del valor del error típico o estándar del error ($s_{y\text{-}y'}$ =1,001), valor necesario para la estimación por intervalo de los valores del criterio. En efecto, para un nuevo sujeto que obtuviera una puntuación igual a 6 en el test de cálculo mental y 60 en el test de razonamiento numérico, la estimación puntual del valor en la nota de matemáticas será:

$$\textit{Nota Matemáticas}' = -1{,}916 + 0{,}709 \cdot 6 + 0{,}052 \cdot 60 = 5{,}458$$

Si realizamos una estimación por intervalo, con un nivel de confianza del 95%:

$$IC^{0{,}95} \rightarrow 5{,}458 \pm 2{,}093 \cdot 1{,}001 = \left[3{,}363 \quad 7{,}553\right]$$

Con una probabilidad de 0,95, el valor en el criterio de un sujeto que obtenga una puntuación 6 en la prueba de cálculo y 60 en el test de razonamiento estará entre 3,363 y 7,553 puntos.

Otras técnicas estadísticas de análisis multivariable

En función del tipo de escala de medida utilizada para las variables criterio y las variables predictoras, será necesario aplicar alguna de las diferentes técnicas de análisis de datos multivariables.

Por ejemplo, si disponemos de una variable criterio medida en escala nominal, y por tanto de tipo cualitativo o categórico, mientras que las variables independientes son de tipo cuantitativo o numérico, podemos utilizar una técnica de clasificación, como el análisis discriminante. En efecto, supongamos

que la variable criterio nota de matemáticas, del ejemplo utilizado, estuviera codificada en suspenso, aprobado, notable, excelente, o simplemente divididos los sujetos entre aprobados y suspendidos.

Aplicar un análisis discriminante nos permitiría determinar la mejor función discriminante, que consiga la clasificación de los sujetos en función de las puntuaciones obtenidas en las pruebas predictoras del cálculo mental y el razonamiento numérico. La función discriminante establecerá la estimación de los pesos (coeficientes) y la combinación lineal de las variables independientes (discriminantes), de modo que los grupos sean, desde el punto de vista estadístico, lo más diferentes posible.

Otra opción puede ser que tanto la variable criterio como las variables independientes se encuentren codificadas en categorías. En este caso, sería aplicable la técnica del modelo *logit*. Este modelo, basado en los modelos lineales logarítmicos, pretende –siguiendo el enfoque de la regresión múltiple– encontrar la expresión de asociación entre la variable criterio y las variables independientes, teniendo en cuenta también la interacción entre las variables independientes.

5.4.3. Validez retrospectiva

La validez concurrente entre uno o varios tests y el criterio, que puede ser útil para la predicción futura de la variable criterio, también en ciertas situaciones puede servir para, dadas ciertas consecuencias medidas a través del criterio, encontrar las causas a los valores obtenidos.

En este caso la variable criterio ha sido registrada anteriormente a las variables predictoras. Por ejemplo, en psicología es habitual la aplicación de diferentes pruebas psicológicas que permitan dar una explicación a determinada conducta de un sujeto.

5.5. Generalización de la validez

El concepto de generalización de la validez se refiere al hecho de extender la validez establecida entre test y criterio a otras situaciones o a grupos de sujetos diferentes a los utilizados inicialmente en el cálculo.

Por ejemplo, imaginemos que se encuentran validadas determinadas pruebas dirigidas a la correcta selección de personal para determinados puestos de administrativo en un banco A. Si es posible utilizar las mismas pruebas para la selección de administrativos en otro banco B, podremos considerar que la validez se ha generalizado. En este caso se trataría de sujetos diferentes a los utilizados inicialmente.

Otro caso podría ser que las pruebas de selección para administrativos en banca se utilicen para la selección de personal administrativo en compañías de seguros. En este caso hablaríamos de generalización también a situaciones diferentes.

6. Evidencia de validez basada en las consecuencias de la aplicación

Cuando se toman decisiones a partir de la aplicación de un test y no se trata solo de describir o interpretar sin que existan acciones que se deriven de ello, se debe pensar en las consecuencias que tiene aplicar dicho cuestionario (Shepard, 1997). Los tests deben usarse cuando se maximicen las consecuencias positivas (beneficios) y se minimicen las negativas (costes) derivadas de su aplicación.

Los tests se aplican esperando que de la información obtenida se extraiga algún tipo de beneficio (poder seleccionar el mejor tratamiento terapéutico, ubicar a los trabajadores de una empresa en el puesto más adecuado, mejorar las técnicas didácticas empleadas, etc.). Uno de los propósitos fundamentales de la validación es indicar en qué casos se pueden obtener estos beneficios.

Dentro de este concepto hay que diferenciar entre evidencias que son relevantes para la validez y evidencias que son importantes para las políticas sociales pero que se sitúan fuera del concepto de validez. Esta diferencia se hace más importante cuando las consecuencias que se derivan del test son diferentes para distintos grupos. Por ejemplo, si se sabe que existen diferencias entre hombres y mujeres en las puntuaciones de un test empleado para la selección de personal, esto va a afectar al uso del test pero no se podría afirmar nada sobre las evidencias de validez basadas en las consecuencias de la aplicación. Para ello, se debe realizar un estudio más pormenorizado de las consecuencias. Si las diferen-

cias se deben a que el aspecto evaluado se distribuye de manera diferente entre los grupos en la población, las diferencias obtenidas no implican que las decisiones que se extraigan de la aplicación del test carezcan de validez. El problema surge cuando dichas diferencias se deben a que se están valorando habilidades que no están relacionadas con la labor que van a realizar los seleccionados o cuando el test es sensible a algunas características de los candidatos que no se pretende que estén relacionadas con el constructo que se va a medir. En el primero de los casos no se puede concluir la falta de indicios de validez respecto a las consecuencias, pero en las dos últimas situaciones sí. No obstante, evidentemente, las tres situaciones serían inadecuadas dentro de las políticas sociales de igualdad de género (APA, 1999).

7. Factores que afectan a la validez

Tal y como se ha comentado anteriormente, uno de los indicios de validez que se pueden (o deben) calcular es la correlación existente entre el test y un criterio ajeno a este. Dicha correlación se puede ver afectada por múltiples factores, como son la fiabilidad de ambas medidas, la longitud del test y la variabilidad (dispersión) de la muestra empleada para obtener las puntuaciones. A continuación se tratarán estos aspectos.

7.1. Fórmulas de atenuación

Cuando se trata de calcular la correlación entre un test y un criterio se parte de las puntuaciones empíricas que se han obtenido en ambos cuestionarios. Dichas puntuaciones están compuestas, según el modelo lineal de Spearman, por la puntuación verdadera y el error de medida, que es aleatorio. Por tanto:

$$X = V_x + e_x$$
$$Y = V_y + e_y$$

X: Puntuación obtenida en el test.
V_x: Puntuación verdadera en el test.

e_x: Error de medida (aleatorio) en el test.
Y: Puntuación obtenida en el criterio.
V_y: Puntuación verdadera en el criterio.
e_y: Error de medida (aleatorio) en el criterio.

Como se puede comprobar, al correlacionar las puntuaciones X e Y también se están correlacionando los dos errores de medida entre sí. Dichos errores son aleatorios, lo que significa que la correlación entre ambos debe ser igual a 0. Por ello, cuanta mayor importancia tengan los errores en la puntuación obtenida (o lo que es lo mismo, cuanto más baja sea la fiabilidad del test y el criterio empleados), menor será la correlación entre X e Y. En definitiva, se puede encontrar una regla según la cual a mayor fiabilidad del test y el criterio, la correlación entre ambos aumentará. Para saber en qué grado lo hace se emplean las fórmulas de atenuación (Spearman, 1907).

7.1.1. Estimación del coeficiente de validez en el supuesto de que el test y el criterio tengan una fiabilidad perfecta

En el caso en el que se supone que el test y el criterio poseen una fiabilidad perfecta se asume que los errores de medida son iguales a 0. En esta situación (en la que el coeficiente de fiabilidad es igual a 1) la puntuación empírica (X o Y según se trate del test o del criterio) es igual a la verdadera.

En este caso la nueva correlación puede calcularse mediante:

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

ρ_{xy} : Coeficiente de validez obtenido al correlacionar las puntuaciones del test y el criterio.
$\rho_{xx'}$: Coeficiente de fiabilidad del test.
$\rho_{yy'}$: Coeficiente de fiabilidad del criterio.

Ejemplo

La correlación entre un test de ansiedad y un criterio (depresión) es de 0,63. La fiabilidad del test es de 0,79 y la del criterio de 0,90. ¿Cuál es la estimación de dicha correlación si se supone que ambos tienen una fiabilidad perfecta?

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} = \frac{0,63}{\sqrt{0,79}\sqrt{0,90}} = 0,75$$

La correlación que se estima entre el test y el criterio si ambos tuviesen una fiabilidad perfecta pasa de 0,63 a 0,75.

7.1.2. Estimación del coeficiente de validez en el supuesto de que el test tenga una fiabilidad perfecta

En el caso de que solo el test tenga una fiabilidad perfecta (igual a 1) la estimación de la nueva correlación viene dada por:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}}$$

En el ejemplo anterior si solo el test tuviese la fiabilidad perfecta, el resultado sería:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}} = \frac{0,63}{\sqrt{0,79}} = 0,71$$

La correlación que se estima entre el test y el criterio al suponer que el test tiene una fiabilidad perfecta cambia de 0,63 a 0,71.

7.1.3. Estimación del coeficiente de validez en el supuesto de que el criterio tenga una fiabilidad perfecta

Si es el criterio el que se supone que tiene una fiabilidad perfecta, la estimación de la correlación viene dada por:

$$\rho_{x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{yy'}}}$$

En el ejemplo anterior:

$$\rho_{xv_y} = \frac{\rho_{xy}}{\sqrt{\rho_{yy'}}} = \frac{0,63}{\sqrt{0,90}} = 0,66$$

La correlación que se estima entre el test y el criterio al suponer que el test tiene una fiabilidad perfecta cambia de 0,63 a 0,66.

7.1.4. Estimación del coeficiente de validez en el supuesto de que se ha mejorado tanto la fiabilidad del test como la del criterio

La situación más frecuente es en la que se mejora la fiabilidad del test, la del criterio o la de ambos pero sin llegar a 1 (este suceso es más teórico que práctico). A continuación se verá cómo estimar la correlación entre test y criterio cuando se mejoran las fiabilidades de ambos. Posteriormente se tratarán las otras opciones.

$$\rho_{XY} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

ρ_{xx} : Es la fiabilidad mejorada del test.

ρ_{yy} : Es la fiabilidad mejorada del criterio.

Ejemplo

La correlación entre un test de ansiedad y un criterio (depresión) es de 0,63. La fiabilidad del test es de 0,79 y la del criterio de 0,90. Añadiendo ítems se consigue incrementar la fiabilidad del test hasta 0,83 y la del criterio hasta 0,92 ¿Cuál es la estimación de dicha correlación tras haber mejorado la fiabilidad de ambos?

$$\rho_{XY} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} = \frac{0,63\sqrt{0,83}\sqrt{0,92}}{\sqrt{0,79}\sqrt{0,90}} = 0,65$$

La estimación de la correlación entre el test y el criterio al haber mejorado la fiabilidad de ambos pasa de 0,63 a 0,65.

7.1.5. Estimación del coeficiente de validez en el supuesto de que se ha mejorado la fiabilidad del test

Cuando solo se mejora la fiabilidad del test, la estimación del nuevo coeficiente de correlación viene dado por:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}}$$

En el ejemplo anterior, si solo se mejora la fiabilidad del test:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}} = \frac{0,63\sqrt{0,83}}{\sqrt{0,79}} = 0,645$$

La estimación de la correlación entre el test y el criterio al haber mejorado la fiabilidad del test pasa de 0,63 a 0,645.

7.1.6. Estimación del coeficiente de validez en el supuesto de que se ha mejorado la fiabilidad del criterio

La estimación del coeficiente de correlación, cuando solo se mejora la fiabilidad del criterio, viene dada por:

$$\rho_{xY} = \frac{\rho_{xy}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}}$$

En el ejemplo anterior, si solo se mejora la fiabilidad del criterio la respuesta sería:

$$\rho_{xY} = \frac{\rho_{xy}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}} = \frac{0.63\sqrt{0,92}}{\sqrt{0,90}} = 0,637$$

La estimación de la correlación entre el test y el criterio al haber mejorado la fiabilidad del criterio pasa de 0,63 a 0,637.

7.1.7. Valor máximo que puede alcanzar el coeficiente de correlación entre test y criterio

Como se puede apreciar, a medida que incrementamos el coeficiente de correlación del test, del criterio o de ambos, el coeficiente de correlación aumenta. Esto solo ocurre hasta cierto punto, ya que el coeficiente de correlación entre test y criterio siempre es menor o igual que su índice de fiabilidad $\left(\rho_{xv} = \sqrt{\rho_{xx'}}\right)$.

Matemáticamente puede representarse como:

$$\rho_{xy} \leq \rho_{xv}$$

Por tanto, en el ejemplo anterior, el máximo coeficiente de correlación que se puede obtener entre el test y el criterio es:

$$\rho_{xv} = \sqrt{\rho_{xx'}} = \sqrt{0,79} = 0,89$$
$$\rho_{xy} \leq 0,89$$

Así pues, el máximo valor del coeficiente de correlación que se puede obtener en ese test es de 0,89.

7.2. Efecto de la longitud del test sobre el coeficiente de correlación test-criterio

Uno de los medios por el que se puede incrementar el coeficiente de correlación test-criterio es aumentando el número de ítems que componen el test. La relación entre el número de ítems y dicha correlación es directa, es decir, a medida que se incremente el número de ítems la correlación aumentará (y se reducirá si se quitan ítems).

La relación entre ambos viene dada por:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1+(n-1)\rho_{xx'}}}$$

ρ_{xy} : Es el valor inicial de la correlación test-criterio.

$\rho_{x'}$: Es el coeficiente de fiabilidad del test.

n: Es el número de veces que se aumenta el test.

Ejemplo

La correlación entre un test de ansiedad de 20 ítems y un criterio (depresión) es de 0,63. La fiabilidad del test es de 0,79. Estima el valor de la correlación test-criterio si se añaden 10 ítems más.

$$n = \frac{20+10}{20} = 1,5$$

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1+(n-1)\rho_{xx'}}} = \frac{0,63\sqrt{1,5}}{\sqrt{1+(1,5-1)0,79}} = 0,65$$

Al añadir 10 ítems a los 20 originales, el coeficiente de correlación incrementa desde 0,63 hasta 0,65.

Otra posibilidad es el hecho de querer llegar hasta un coeficiente de correlación que se desee y por tanto haya que calcular el número de ítems que se deben añadir al cuestionario para poder alcanzarlo. Para ello:

$$n = \frac{(1-\rho_{xx'})\rho^2_{Xy}}{\rho^2_{xy} - \rho^2_{Xy}\rho_{xx'}}$$

ρ^2_{Xy} : Es el cuadrado del coeficiente de correlación deseado.

Ejemplo

La correlación entre un test de ansiedad de 20 ítems y un criterio (depresión) es de 0,63. La fiabilidad del test es de 0,79. ¿Cuántos ítems habrá que añadir al test si se pretende alcanzar un coeficiente de correlación test-criterio de 0,67?

$$n = \frac{(1-\rho_{xx'})\rho^2_{Xy}}{\rho^2_{xy} - \rho^2_{Xy}\rho_{xx'}} = \frac{(1-0,79)0,67^2}{0,63^2 - 0,67^2 0,79} = 2,23$$

El test debería ser incrementado 2,23 veces. Dado que el test original tiene 20 ítems: $20 \times 2,23 = 44,6$. Evidentemente no se pueden tener decimales (estamos hablando de ítems, no se puede tener una porción de ítem en el test), por lo que lo debemos ajustar. Cuando estemos añadiendo ítems, el ajuste siempre se debe hacer hacia el entero superior, es decir, en este caso tendría que haber 45 ítems. En el caso de que el test sea excesivamente largo y no nos importe eliminar ítems hasta llegar a un coeficiente de correlación test-criterio menor (pasar de 0,63 a 0,50 por ejemplo), el ajuste se deberá hacer al entero inferior.

7.3. Efecto de la variabilidad de la muestra en la correlación test-criterio

El coeficiente de correlación se ve muy afectado por la dispersión de la muestra en la que esté calculado. La relación entre la dispersión y la correlación es directa: a mayor dispersión se obtendrá una mayor correlación.

En algunos campos de la psicología es muy frecuente que solo se pueda calcular la correlación entre el test y el criterio en una pequeña muestra de perso-

nas. El ejemplo más claro es el de la selección de personal. Tras haber empleado un test para seleccionar de entre los candidatos a los más adecuados al puesto, solo se puede correlacionar la puntuación obtenida en el test con un criterio como el de rendimiento laboral. En este caso, la dispersión de los seleccionados será menor que la del total de candidatos, ya que se selecciona a las personas que tienen características muy similares (y que más se ajustan a las buscadas para el puesto).

Tras calcular el coeficiente de correlación en la muestra de seleccionados puede interesar estimar cuál sería si se hubiese calculado sobre la totalidad de aspirantes. Para ello, se debe partir de dos supuestos: la pendiente de la recta de regresión es la misma para los dos grupos (admitidos y aspirantes) y el error típico de estimación también es el mismo para ambos grupos. Matemáticamente:

$$a)\frac{\rho_{xy}\sigma_y}{\sigma_x} = \frac{\rho_{XY}\sigma_Y}{\sigma_X}$$

$$b)\sigma_y\sqrt{1-\rho_{xy}^2} = \sigma_Y\sqrt{1-\rho_{XY}^2}$$

Donde las letras mayúsculas hacen referencia al grupo de admitidos y las minúsculas al total de aspirantes.

Para estimar el valor de la correlación test-criterio en el total de aspirantes tras haberla calculado en el de admitidos solo es necesario aplicar:

$$\rho_{xy} = \frac{\sigma_x\rho_{XY}}{\sqrt{\sigma_x^2\rho_{XY}^2 + \sigma_X^2 - \sigma_X^2\rho_{XY}^2}}$$

ρ_{XY} : Coeficiente de correlación test-criterio en la muestra de admitidos.

σ_x^2 : Varianza en el test del total de aspirantes.

σ_X^2 : Varianza en el test de los admitidos.

Ejemplo

Se aplicó un test de asertividad a 1.000 personas para seleccionar a 10 sobrecargos de vuelo. La desviación típica obtenida en el test por el total de aspirantes fue de 15 y en la muestra de admitidos de 4. Tras un tiempo trabajando se comprobó que la corre-

lación entre las puntuaciones en el test y la ejecución laboral (valorada por sus superiores) fue de 0,36. ¿Cuál sería el coeficiente de correlación test-criterio si se hubiese calculado sobre el total de los aspirantes?

$$\rho_{xy} = \frac{\sigma_x \rho_{XY}}{\sqrt{\sigma_x^2 \rho_{XY}^2 + \sigma_X^2 - \sigma_X^2 \rho_{XY}^2}} = \frac{15 \times 0,36}{\sqrt{15^2 \times 0,36^2 + 4^2 - (4^2 \times 0,36^2)}} = 0,82$$

Como se puede observar, hay un incremento notable en el valor del coeficiente de correlación debido a la diferente dispersión que tienen los grupos.

Capítulo IV

Transformación e interpretación de las puntuaciones

Sergi Valero

Interpretar el valor obtenido por una persona en una medida psicológica, y hacerlo de un modo razonable y razonado, exige no solo conocer con cierto detalle las características técnicas del instrumento que es empleado, sino también conocer el constructo al que se hace referencia y entender cómo este se estructura y se explica en un marco teórico que tiene sentido, a la vez, en un marco disciplinario concreto, ya sea educacional, clínico o de cualquier otro tipo.

Es evidente que un capítulo como este no puede alcanzar satisfactoriamente un objetivo de esta magnitud. Por esta razón, cuando en este texto se hable de *interpretar una puntuación*, se deberá siempre considerar desde una perspectiva eminentemente metodológica y matemática y, lo más importante, inferida de acuerdo con la respuesta de un conjunto de individuos que, evaluados con una misma medida, proporcionarán el modelo de comportamiento que será dispuesto a modo de norma, de marco de referencia respecto al cual la puntuación de una persona concreta será interpretada. Bajo esta perspectiva tendrán su sentido las puntuaciones transformadas. Los percentiles, las puntuaciones estandarizadas y las normas cronológicas se recogerán aquí, exponiendo sus características más importantes, la manera de calcularlas y, donde se considere oportuno, sus limitaciones más relevantes.

Un segundo término de gran relevancia, y que será presentado en el tercer apartado, es el concepto de baremo. Se trata de una herramienta de gran importancia en la que se establece, de manera estructurada, una conexión entre la puntuación de un individuo y el comportamiento normativo proporcionado por una muestra relevante de sujetos que, en algunos casos, aparecerán estratificados según conveniencia. Se dedicará un especial interés a la determinación

de cuál debe ser esta muestra normativa. En este sentido, se expondrán los criterios de relevancia, representatividad y homogeneidad como factores fundamentales en la determinación de la calidad de esta muestra de referencia.

El capítulo finalizará exponiendo las diferentes estrategias disponibles para establecer equiparaciones entre puntuaciones de instrumentos diferentes que evalúan un mismo rasgo. Un conjunto de estrategias, también metodológicas y matemáticas, que son especialmente interesantes en aquellos ámbitos de la evaluación en los que los procesos de memoria o aprendizaje de las personas participantes pueden convertirse en una limitación relevante de las inferencias que se derivan de ellos.

1. Interpretación de una puntuación

Consideremos una circunstancia nada excepcional en la que para aprobar una asignatura, por ejemplo *Lengua inglesa*, es necesario obtener una puntuación de 5, asumiendo que estamos ante un examen clásico en el que la puntuación mínima posible es 0 y la máxima 10. ¿Alguna vez os habéis preguntado qué propiedad intrínseca tiene el número 5 como para considerar que con él y por encima de él uno ya domina esta lengua, y que por el contrario, cuando se está debajo, nuestras aptitudes no satisfacen el nivel exigido? Excepto que nos sintiéramos confortables con algún postulado más propio de la numerología que de la psicometría, habría que considerar que, *a priori*, el número 5 carece de cualquier atributo intrínseco que lo haga especial. Y lo mismo se podría afirmar de una puntuación de 4 o de cualquier otra. Si no se nos informa de nada más, deberemos asumir que aprobar o suspender en función de superar o no una puntuación de 5 responde a una simple convención. Hay que establecer un punto de corte porque muchas veces es necesario tomar una decisión respecto a quién supera un curso, y el valor de 5, cifra que ocupa la posición central, suele ser el valor de referencia utilizado más frecuentemente.

En otras circunstancias, sin embargo, el criterio empleado para discernir entre personas con o sin un determinado rasgo no responde a un juicio arbitrario o convencional. El criterio puede responder a consideraciones empíricamente

contrastadas, surgido generalmente de la comparación de poblaciones de sujetos muy diferenciados en el rasgo objeto de interés y que permiten la obtención de un punto de corte, de una puntuación de referencia dotada de capacidad discriminativa (Ramos-Quiroga et al., 2009).

En muchas situaciones, no obstante, no se dispone de criterios externos que proporcionen un marco de interpretación respecto al que otorgar relevancia o significación a una puntuación obtenida en una medida. De hecho, se puede afirmar que la inmensa mayoría de las medidas psicológicas carecen de criterios externos e independientes con los que inferir el estatus actitudinal, aptitudinal o clínico de un individuo. ¿Qué se puede inferir entonces de la puntuación de una persona en una medida cuando no están disponibles estos referentes?

Mirar *dentro* del instrumento, es decir, remitirse a alguna de las propiedades estructurales de la medida es una alternativa posible. Una aproximación sencilla, por ejemplo, consiste en atender las puntuaciones mínima y máxima que es posible alcanzar. Esta aproximación absoluta permite contextualizar una puntuación indicando su excepcionalidad. Cualquier manual de uso o publicación de referencia de una medida debería proporcionar información clara respecto a estos valores mínimo y máximo. La sencillez evidente de esta aproximación, no obstante, suele ser a la vez su limitación más importante: probablemente tendremos menos problemas para emitir un juicio sobre una persona que obtiene en una media una puntuación de 17 que para una de 14, sabiendo que la puntuación máxima alcanzable es 18. Ahora bien, ¿cuánto debe estar un valor alejado del valor máximo (o mínimo) posible para considerar que es mucho o poco? No es posible dar una respuesta operativa y universal a esta pregunta.

Una manera alternativa de proceder consiste en comparar la puntuación obtenida por una persona con la puntuación que obtienen otras personas que fueron evaluadas con la misma medida. La media (y también la mediana) suele ser una manera frecuente de resumir el comportamiento de este grupo de personas. De este modo, si una persona obtiene en una medida de aptitudes mnésicas una puntuación de 7 y la media de sus colegas fue de 9, evaluados siguiendo la misma pauta, podemos afirmar que esta persona presenta una capacidad de memoria inferior a la presentada por sus compañeros. Y aquí descansa una de las claves de esta aproximación relativa: la inferencia que se hace de una observación concreta es una función de la tendencia de comportamiento de otras personas en la misma medida. Retomaremos esta observación más adelante.

2. Transformación de las puntuaciones

Según lo que se ha apuntado hasta el momento, es posible disponer de estrategias que permiten la interpretación de una medida y que, además, no suponen ninguna manipulación de los datos obtenidos. No obstante, no son las únicas estrategias al alcance para dar sentido a una determinada puntuación. Otros procedimientos, menos simples que los expuestos hasta el momento, ofrecen al usuario una aproximación más precisa e informativa. Estos procedimientos implican, no obstante, la transformación de las puntuaciones observadas.

Transformar las puntuaciones de una medida consiste en aplicar una estrategia de codificación en la que las puntuaciones obtenidas en la medida, sus puntuaciones directas, son recodificadas (transformadas) en un nuevo sistema de valores que facilitan al usuario su interpretación.

Una propiedad necesaria de las puntuaciones transformadas es que no alteran el escalamiento de las puntuaciones directas, es decir, respetan la diferente disposición de los distintos sujetos según la magnitud de sus puntuaciones. Y este mantenimiento de la ordenación original se lleva a cabo, como se ha mencionado, otorgando a las nuevas puntuaciones un sentido eminentemente práctico. Los percentiles, las puntuaciones estandarizadas, las puntuaciones estandarizadas derivadas y normalizadas y las normas cronológicas serán las distintas estrategias de transformación que serán tratadas a continuación.

2.1. Percentiles

Para describir el concepto de percentil, también conocidos como centiles, se expondrá en primer lugar el concepto de mediana. Una vez ordenados los valores de un conjunto de observaciones, la mediana será aquel valor que ocupa la posición u orden central. Un ejemplo sencillo ayudará en esta primera aproximación.

Ejemplo

Se dispone del siguiente conjunto de valores: 4, 5, 4, 3, 7, 8, 3, 1 y 6. En primer lugar habrá que ordenar la secuencia: 1, 3, 3, 4, 4, 5, 6, 7, 8. La mediana de este conjunto de valores es el 4 (el segundo 4), dado que es el valor que ocupa la posición central de la distribución. La mitad de las personas están dispuestas a un lado y otro de este valor.

En este ejemplo ha sido fácil identificar el valor central de la distribución, ya que el número de observaciones es impar. Toda distribución impar tiene siempre un valor que ocupa la posición central. ¿Y si la distribución de valores fuese par? Añadimos a la ya conocida y ordenada distribución de valores un nuevo valor, cualquiera: 1, 3, 3, 4, 4, 5, 6, 7, 8 y 9. En una distribución con número par de observaciones no es uno, sino dos, los números que se ubican en posición central. En el supuesto que nos ocupa son los valores 4 y 5. Aplicando una media aritmética entre los dos valores se concluirá que la mediana de esta nueva distribución de valores es 4,5. Todo es correcto: una mediana que proviene de valores enteros puede ser decimal y, además, no tiene por qué coincidir con ninguna de las observaciones originales.

¿Qué tienen que ver, no obstante, los percentiles con la mediana? Los percentiles son una generalización de la mediana. La transformación basada en los percentiles consiste en asignar a cada puntuación directa una puntuación porcentual, según la posición de las observaciones dentro del conjunto de observaciones. Si el valor porcentual es del 25%, se estará hablando del percentil 25. Y bajo esta denominación se estará identificando aquella puntuación directa que deja por debajo una cuarta parte de todas las observaciones. O expresado de otra manera, aquel valor que con él y por encima de él se encuentra el 75% de todas las observaciones. Si el porcentaje asciende a un 50%, se hablará del percentil 50, que se corresponderá con el valor que ocupará la posición central de la distribución. El percentil 50 es, efectivamente, la mediana.

La primera formulación práctica de los percentiles se remonta a finales del siglo XIX, y proviene del explorador y científico Francis Galton, primo de Charles Darwin, que entre otras aficiones estaba interesado en la medida de multitud de rasgos antropométricos. Hoy en día, y sin abandonar el ámbito de la medida de las propiedades del cuerpo, los percentiles continúan siendo una estrategia de interés, por ejemplo, para aquellos padres y madres que desean conocer en qué grado sus hijos se ajustan a los otros niños de la misma edad en términos de altura y peso.

Los percentiles han resultado una medida muy empleada no solo en el contexto de la medida del peso o de la altura, sino también en el contexto de las medidas psicológicas. Su simplicidad y universalidad, dado que no hay que tener amplios conocimientos en matemáticas ni de estadística en particular, ha estimulado sin duda su difusión como estrategia interpretativa. Expongamos un caso práctico que servirá para ejemplificar el sentido y generalizar su cálculo.

En una muestra de 1.000 personas de nivel académico, edad y género diferentes (Gomà-i-Freixament et al., 2008) fue administrado el Zuckerman-Kulhman Personality Questionnaire (ZKPQ, Zuckerman, Kuhlman, Joireman, Teta y Kraft, 1993), un instrumento que mide cinco dimensiones básicas de personalidad normal.

De las cinco dimensiones que el modelo permite medir, se ha centrado la atención, por motivos de simplificación, en una de las variables de personalidad, el neuroticismo[1]. En este instrumento, esta dimensión de personalidad es operativizada mediante un total de 19 ítems de respuesta tipo verdadero-falso, donde el sumatorio de todos ellos (teniendo en cuenta que algunos ítems son directos y otros inversos) resulta ser una estimación del grado en el que un individuo tiende a presentar este rasgo de personalidad. Hay que estar atento a las peculiaridades de cada instrumento para saber cómo debe ser obtenida una puntuación en la variable de interés. No siempre se sigue una misma pauta para calcular las puntuaciones totales.

En el supuesto que nos ocupa, la puntuación máxima alcanzable es 19 y la mínima 0. Una mayor puntuación equivale a una mayor intensidad del rasgo. La distribución de valores absolutos y relativos de este grupo de 1.000 personas en esta variable aparece recogida en la tabla 1.

Tabla 1. Puntuaciones directas, frecuencias, porcentajes y percentiles

Puntuación en la escala	Frecuencia absoluta (fi)	Frecuencia acumulada (fa)	Porcentaje (%)	Porcentaje acumulado (%)	Percentil (Pc)
0	21	21	2,1	2,1	0
1	55	76	5,5	7,6	3
2	49	125	4,9	12,5	8
3	77	202	7,7	20,2	13

1. El neuroticismo expresa la tendencia de toda persona a mostrarse alterada, con tensión, preocupada, con indecisión y con labilidad emocional (estado de humor frecuentemente cambiante).

Puntuación en la escala	Frecuencia absoluta (fi)	Frecuencia acumulada (fa)	Porcentaje (%)	Porcentaje acumulado (%)	Percentil (Pc)
4	63	265	6,3	26,5	21
5	70	335	7,0	33,5	27
6	76	411	7,6	41,1	34
7	67	478	6,7	47,8	42
8	68	546	6,8	54,6	48
9	72	618	7,2	61,8	55
10	59	677	5,9	67,7	62
11	60	737	6,0	73,7	68
12	57	794	5,7	79,4	74
13	44	838	4,4	83,8	80
14	43	881	4,3	88,1	84
15	41	922	4,1	92,2	89
16	38	960	3,8	96,0	93
17	23	983	2,3	98,3	97
18	11	994	1,1	99,4	99

Si se toma por ejemplo la puntuación 8, se puede observar que han sido 68 las personas que han presentado esta puntuación (columna de frecuencia absoluta). Dado que son un total de 1.000 personas en esta muestra, se puede afirmar que un 6,8% de los participantes tienen una puntuación directa de 8 (columna de porcentaje). La penúltima columna, identificada como porcentaje acumulado, ha exigido que los valores directos fueran dispuestos de manera ordenada. Los porcentajes correspondientes a cada puntuación se van sumando y acumulando. En el caso de la puntuación directa 8, el porcentaje acumulado es 54,6. Esto significa que el 54,6% de las personas de la muestra han obtenido puntuaciones de como máximo 8 unidades (8 o menos).

¿Cuál es el percentil asociado a este valor directo? Dado que debe ser el porcentaje de personas que quedan por debajo de este valor, habrá que observar que hasta llegar a una puntuación directa de 7 se observa un porcentaje acumulado del 47,8%. Este valor, una vez redondeado al entero inmediatamente superior, pasa a ser de un 48%. La puntuación directa 8 se asocia a un percentil de 48.

Si la correspondencia entre puntuaciones directas y percentiles se enfocara de manera inversa, en cambio, se podría formular una pregunta del tipo "¿A qué valor directo correspondería un percentil de 55?". Según los porcentajes acumulados, hasta un valor directo de 8 le corresponde un porcentaje acumulado de 54,6%. Un porcentaje de 55, a pesar de que por poco, no se ha producido todavía. El siguiente valor directo es el 9, y hasta él se ha acumulado el 61,8% de las observaciones. En este caso sí, el 55% objeto de interés ya se ha producido, dado que está incluido en el 61,8%. La puntuación directa 9, por lo tanto, se asocia a un percentil de 55.

De acuerdo con este razonamiento habrá que entender por qué una puntuación de 9 es también percentil 56. Pero también 57, 58, 59, 60 y 61. Solo cuando se llega a la puntuación directa 10 el percentil pasa a ser 62. Esta circunstancia, en la que un mismo valor directo se vincula a más de un percentil, es del todo plausible cuando la variable objeto de interés no dispone de valores suficientes para cubrir las cien unidades para las que el percentil sí está preparado.

Hay que tener en cuenta que esta estrategia para calcular percentiles, basada en los porcentajes acumulados, no es la única posible. Existen otras aproximaciones que son también aceptadas. Y decimos aproximaciones porque el estudiante deberá tolerar que no todas ellas lleguen necesariamente a idéntico resultado. Una estrategia alternativa a la ya expuesta para calcular un percentil, cualquiera, es la que se propone a continuación:

$$P = \frac{fa + 0,5f_y}{N} \times 100$$

f_a: Frecuencia acumulada previa a la puntuación directa de la que se quiere calcular el percentil.

f_y: Frecuencia absoluta en la que se encuentra la puntuación directa.

N: Número de personas que constituyen la muestra.

Así pues, y de acuerdo con la tabla 1, el percentil que se asocia a una puntuación directa de 9 es 58:

$$P9 = \frac{546 + (0,5 \times 72)}{1000} \times 100 \approx 58$$

Dado que los percentiles no tienen decimales, se ha ignorado el decimal del resultado exacto obtenido (58,2). Hay que recordar que siguiendo la primera estrategia expuesta para calcular percentiles, la puntuación directa 9 era, a la vez, los percentiles 55, 56, 57, 58, 59, 60 y 61. El valor obtenido en la fórmula anterior es solo uno de los valores, el central[2].

La primera de las ventajas del uso de los percentiles, como ya se ha indicado, es su simplicidad conceptual. Una segunda ventaja es que en el cálculo y la correcta interpretación de este, dado que centra su atención solo en la orden en el que las observaciones son dispuestas, resulta irrelevante el modo como se distribuyen los valores objeto de análisis. Esto implica que, a pesar de que hay estrategias de transformación que requieren una distribución de valores esencialmente ajustada a una curva normal, más adelante se verá, en el caso de los percentiles es posible abordar cualquier tipo de distribución. Esta característica es especialmente interesante en psicología, donde muchas de las variables de interés raramente se ajustan a una distribución normal de valores.

Como se ha visto, los percentiles consisten en dividir una distribución de valores en cien posiciones ordinales. No obstante, existen otros abordajes que simplifican la cantidad de órdenes posibles y que pueden ser suficientes para resumir el comportamiento de un conjunto de valores. Los más comunes son:

- Los *cuartiles*. Dividen una distribución en cuatro partes. Se suelen identificar como cuartil 1, cuartil 2 y cuartil 3. Coinciden con los percentiles 25, 50 y 75, respectivamente.
- Los *quintiles*, que al dividir la distribución en cinco partes se equiparan a los percentiles 20, 40, 60 y 80.

2. En Wikipedia (en inglés) y en otros recursos de Internet, encontraréis otras estrategias alternativas, también válidas, para calcular los percentiles.

– Los *deciles*. Suponen una división en diez partes y se corresponden con los percentiles 10, 20, 30, 40, etc.

2.2. Puntuaciones estandarizadas

Si bien la media es un estadístico de tendencia central que permite expresar el comportamiento medio de un conjunto de observaciones, la desviación típica (o estándar) permite cuantificar lo diferentes que son entre sí estas observaciones. Este índice de dispersión expresa la discrepancia promedio entre una observación cualquiera y la media del conjunto de observaciones. La media y la desviación típica son la materia prima imprescindible para calcular una puntuación estandarizada. Si bien los percentiles dirigían la atención hacia la posición que ocupa un individuo respecto a un grupo de referencia, en el caso de las puntuaciones estandarizadas el interés se focaliza en la discrepancia que presenta este individuo respecto al modo de comportamiento promedio del grupo, su media.

La puntuación estandarizada o puntuación típica es aquella que permite expresar cuántas desviaciones típicas por encima o por debajo de una media se sitúa una observación. Este proceso de estandarización tiene su sentido cuando la distribución de valores objeto de interés se ajusta esencialmente a una curva normal. El valor resultante de esta transformación se constituye en una nueva puntuación, que recibe el nombre de puntuación estandarizada z. Su cálculo responde a la fórmula siguiente:

$$Zx = \frac{X - \overline{X}}{Sx}$$

X: Puntuación directa.

$\overline{X}$: Media de la muestra.

S_x: Desviación típica de la muestra.

Hay que tener presente que esta puntuación *z* tendrá siempre una media de 0 y una desviación estándar de 1. En la tabla 2, y retomando el estudio sobre neuroticismo en una muestra de 1.000 personas:

Tabla 2. Puntuaciones directas, frecuencias y puntuaciones estandarizadas *Z*

Puntuaciones directas	Frecuencia absoluta	Puntuaciones estandarizadas z
0	21	-1,72
1	55	-1,51
2	49	-1,30
3	77	-1,09
4	63	-0,88
5	70	-0,66
6	76	-0,45
7	67	-0,24
8	68	-0,03
9	72	0,18
10	59	0,39
11	60	0,61
12	57	0,81
13	44	1,03
14	43	1,24
15	41	1,45
16	38	1,67
17	23	1,88
18	11	2,09
19	6	2,30

Para poder calcular los valores z, ubicados en la última columna de la tabla, es necesario conocer la media y la desviación típica de la variable en esta muestra: la media era 8,14 y la desviación típica 4,72. Así pues, en el caso con-

creto de una persona que ha obtenido una puntuación directa de 8, su puntuación z será:

$$-0,03=\frac{8-8,14}{4,72}$$

¿Qué significa este valor de –0,03? Dado que la media de z siempre es 0 y que su desviación típica también es siempre 1, observar una puntuación típica de –0,03 significa que los sujetos que tienen una puntuación directa de 8, y que son un total de 68 personas, se ubican en 0,03 desviaciones típicas de la media de la muestra. Sabemos, además, que este valor directo se ubica por debajo de la media porque su puntuación típica es negativa. Necesariamente, todos los sujetos que presentan puntuaciones directas que se disponen por debajo de la media obtendrán puntuaciones z negativas, mientras que aquellos que se ubican por encima de la media presentan puntuaciones típicas positivas. Si nos fijamos en la parte final de la tabla 2, veremos que las 6 personas que obtienen una puntuación de 19 tienen una puntuación típica de 2,30, lo que implica que estas se ubican a más de 2 desviaciones típicas por encima de la media. En cambio, las 21 personas que obtuvieron una puntuación directa de 0 presentan una puntuación típica de –1,72. Son personas que quedan, por lo tanto, a 1,72 desviaciones por debajo de la media.

Una pregunta importante que nos podemos formular es si distanciarse en 1,72 o 2,30 desviaciones típicas de una media es mucho o poco. Está claro que alejarse de una media en 2,30 desviaciones típicas es hacerlo más que si lo hacemos 1,72 veces. Podéis ver en la tabla 2 que en esta muestra las personas con la máxima puntuación se sitúan más alejadas de la media que las personas que obtienen la puntuación más baja posible. Ahora bien, ¿es mucho hacerlo 2,3 veces?

En el estudio de Valero et al. (2012), se comparan dos muestras de sujetos adultos: una constituida por 217 personas adultas afectadas por un trastorno por déficit de atención e hiperactividad (TDAH) y la otra formada por 434 sujetos control emparejados por edad y género. Distintas variables de personalidad fueron contrastadas, pero para simplificar la exposición se centrará de nuevo la atención sobre la dimensión de neuroticismo. La tabla 3 expone las puntuaciones típicas en esta variable para cada uno de los dos grupos de estudio.

Tabla 3. Puntuaciones directas y estandarizadas para una muestra de personas con TDAH y controles

Puntuaciones directas	**Puntuación estandarizada z**	
	TDAH	Controles
0	-2,54	-1,67
1	-2,38	-1,46
2	-2,16	-1,25
3	-1,94	-1,04
4	-1,73	-0,83
5	-1,51	-0,62
6	-1,30	-0,41
7	-1,08	-0,19
8	-0,87	0,02
9	-0,65	0,23
10	-0,43	0,44
11	-0,22	0,65
12	0	0,86
13	0,21	1,08
14	0,43	1,29
15	0,65	1,50
16	0,86	1,71
17	1,08	1,92
18	1,29	2,14
19	1,51	2,35
	Media = 12,01	Media = 7,91
	Desviación típica = 4,63	Desviación típica = 4,72

Como se puede ver a pie de tabla, los sujetos afectados por TDAH presentan una media en la medida de neuroticismo superior a la presentada por los sujetos control (12,01 frente a 7,91). Se trata de un resultado esperable, teniendo en cuenta que una es muestra psiquiátrica y la otra no. Las respectivas desviaciones típicas, en cambio, y en este estudio, son muy comparables. Si para cada puntuación directa se comparan las respectivas puntuaciones típicas de las dos muestras, se observará que no coinciden en ningún caso. De hecho, se muestran claramente diferenciadas.

En el caso de las personas afectadas por TDAH, las puntuaciones tipificadas tienden a estar sistemáticamente desplazadas hacia la parte baja de la distribución, respecto a las presentadas por los sujetos control. La discrepancia entre las dos medias condiciona este comportamiento. Si bien una puntuación directa de 0 en el caso de los sujetos control implica ubicarse en 1,67 desviaciones típicas por debajo de la media, en el caso de los pacientes afectados de TDAH, la misma puntuación directa los aleja más de dos desviaciones típicas y media de la respectiva media. Todo depende del grupo en el que se esté ubicado.

Y aquí descansa una de las consideraciones más relevantes que hay que tener presente cuando se interpretan puntuaciones estandarizadas: las inferencias a las que podemos llegar mediante el uso de estas puntuaciones dependen de la muestra que es empleada, dado que es de ella de donde se obtienen la media y desviación típica que permiten el cálculo de las puntuaciones transformadas. Elegir cuidadosamente la muestra de referencia resulta imprescindible. Más adelante, cuando se hable del concepto de baremos, se retomará esta idea.

2.3. Puntuaciones estandarizadas derivadas

Aunque intrínsecamente no es un problema, las puntuaciones *z* se expresan con valores negativos y decimales. El sentido básico de las puntuaciones estandarizadas derivadas es transformar las puntuaciones típicas según un sistema de codificación que elimina estos valores negativos y decimales. Una de las pun-

tuaciones típicas derivadas más comúnmente empleadas es la puntuación *T* de McCall. Su formulación es sencilla:

$$T = 50 + 10z$$

Consiste en partir de una constante, en este caso 50, y sumarle diez veces el valor de la puntuación típica *z*. Para interpretar debidamente esta puntuación solo hay que tener presente que la media de esta puntuación es 50 y su desviación típica 10. De este modo, por ejemplo, una puntuación *z* de 2 tendrá un valor de 70 una vez transformada en puntuación *T*. En ambos casos (*z* y *T*), y necesariamente, sería indicador de una puntuación directa que se ubica por encima de la correspondiente media en dos desviaciones típicas.

La tabla 4 recoge los valores de la tabla 2 añadiendo las puntuaciones *T* correspondientes:

Tabla 4. Puntuaciones directas, frecuencias y puntuaciones estandarizadas *z* y *T*

Puntuaciones en la escala	Frecuencia absoluta	Puntuación estandarizada *z*	Puntuación estandarizada derivada *T*
0	21	-1,72	33
1	55	-1,51	35
2	49	-1,30	37
3	77	-1,09	39
4	63	-0,88	41
5	70	-0,66	43
6	76	-0,45	46
7	67	-0,24	48
8	68	-0,03	50
9	72	0,18	52
10	59	0,39	54
11	60	0,61	56

Puntuaciónes en la escala	Frecuencia absoluta	Puntuación estandarizada *z*	Puntuación estandarizada derivada *T*
12	57	0,81	58
13	44	1,03	60
14	43	1,24	62
15	41	1,45	65
16	38	1,67	67
17	23	1,88	69
18	11	2,09	71
19	6	2,30	73

Es posible encontrar puntuaciones típicas derivadas de distinta naturaleza, y es que estrictamente cualquiera puede crearse la suya, dado que una puntuación derivada se formula según criterios arbitrarios. Ejemplos diferentes los encontraremos en el test de inteligencia Standford-Binet, que emplea una media de 100 y una desviación típica de 16, la Wechler Adult Intelligence Scale (WAIS), con una media también de 100, pero con una desviación típica de 15, o el Minnesota Multiphasic Personality Inventory (MMPI), que, coincidiendo con la escala de McCall, se interpreta en torno a un media de 50 y una desviación típica de 10.

2.4. Puntuaciones estandarizadas normalizadas

Si se asume que la distribución de valores que es objeto de análisis se ajusta a una curva normal, es posible transformar los percentiles en aquellas puntuaciones estandarizadas que se corresponderían si la distribución fuera, efectivamente, normal. Estas puntuaciones recibirían el nombre de *puntuaciones estandarizadas normalizadas*.

Tabla 5. Puntuaciones directas, frecuencias, puntuaciones estandarizadas z y T y normalizada

Puntuación en la escala	Frecuencia absoluta (f)	Frecuencia acumulada (fa)	Porcentaje acumulado	Percentil (Pe)	Puntuación estandarizada z	Puntuación estandarizada normalizada
0	21	21	2,1	2	-1,72	-2,05
1	55	76	7,6	3	-1,51	-1,88
2	49	125	12,5	8	-1,30	-1,41
3	77	202	20,2	13	-1,09	-1,13
4	63	265	26,5	21	-0,88	-0,81
5	70	335	33,5	27	-0,66	-0,61
6	76	411	41,1	34	-0,45	-0,41
7	67	478	47,8	42	-0,24	-0,20
8	68	546	54,6	48	-0,03	-0,05
9	72	618	61,8	55	0,18	0,13
10	59	677	67,7	62	0,39	0,31
11	60	737	73,7	68	0,61	0,47
12	57	794	79,4	74	0,81	0,64
13	44	838	83,8	80	1,03	0,84
14	43	881	88,1	84	1,24	0,99
15	41	922	92,2	89	1,45	1,23
16	38	960	96,0	93	1,67	1,48
17	23	983	98,3	97	1,88	1,88
18	11	994	99,4	99	2,09	2,33
19	6	1000	100,0	100	2,30	-

De acuerdo con la tabla 5, si se toma por ejemplo la puntuación directa 9, se observará que, como ya se sabía, su puntuación estandarizada *z* es 0,18 y su per-

centil 55. El paso siguiente es buscar una tabla estandarizada de puntuaciones normales, también conocida como tabla de puntuaciones *z* (véase la tabla al final de este manual). La tabla 6 es el extracto de una de estas tablas. Si en ella se busca una proporción de 0,55 se localizará el valor *z* correspondiente.

En el supuesto que nos ocupa la proporción más cercana es 0,5517, y en consecuencia el valor z que se vincula a ella es 0,13. Cualquier percentil, a modo de proporción, tiene asociado un valor *z* y viceversa.

Tabla 6. Extracto tabla puntuaciones *z*

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224

Reiteramos que el valor obtenido en una tabla de valores *z* asume que la distribución de valores que es objeto de interés se ajusta básicamente a una curva normal. La discrepancia observada entre las puntuaciones 0,18 y 0,13 (tabla 5), o genéricamente entre cualquier puntuación estandarizada calculada y la correspondiente puntuación estandarizada normalizada, responde a la discrepancia existente entre la distribución de valores empíricos y una distribución teórica normal. Hay que tener presente que la presencia de discrepancia entre las dos distribuciones de valores tipificados no será la excepción, sino la norma. Como ya hemos mencionado con anterioridad, son muchas las variables que no se ajustarán a una (perfecta) distribución normal. Para aquellas que lo hagan esencialmente, he aquí una manera posible de transformar una puntuación.

Del mismo modo que en el caso de las puntuaciones estandarizadas derivadas, es posible disponer también de puntuaciones estandarizadas normalizadas

derivadas. En psicología las más habituales son las estaninas o eneatipos y los decatipos. Las respectivas formulaciones son:

$$\text{Eneatipos} = 5 + 2z_n$$
$$\text{Decatipos} = 5{,}5 + 2z_n$$

Donde la *z* es la puntuación estandarizada normalizada.

Asumiendo una distribución normal de valores, si clicáis en un buscador términos como *z distribution percentile*, podréis ver gráficamente la correspondencia que se establece entre algunos de los procedimientos de transformación presentados en este capítulo.

2.5. Normas cronológicas

Otra posible transformación de las puntuaciones directas consiste en hacerlo como normas cronológicas o de edades. La estrategia a seguir consiste en administrar a varios grupos de edad, representativos de una determinada población, una misma medida, recogiendo para cada grupo de edad la respectiva media. Posteriormente, cuando un individuo es evaluado, su puntuación es transformada en una edad, aquella edad que en la población de referencia se aproximaría a la puntuación obtenida. La idea básica que justifica esta fórmula es simple: asumir que existen formas de comportamiento que tienden a ser típicos en una edad. Así pues, si después de seleccionar una muestra representativa de una determinada población se administra una determinada medida a los niños de 6 años y la media obtenida es de 12, y la misma medida administrada a los niños de 8 años es de 17, si un niño concreto obtiene una puntuación de 17, se afirmará que en esta medida el niño tiene una edad de 8 años (aunque cronológicamente tenga 7).

Uno de los contextos donde ha sido más popular esta estrategia es el del cociente de inteligencia. Su formulación es:

$$\text{QI} = \frac{EM}{EC} \times 100$$

EM: Edad mental.
EC: Edad cronológica.

La transformación en edades es una estrategia antigua, probablemente tanto como lo ha sido la medida de los constructos psicológicos. No obstante, no se puede dejar de mencionar que esta transformación se acompaña de limitaciones nada desdeñables, lo que han ido reduciendo a lo largo del tiempo su presencia en la bibliografía. Una de sus limitaciones más relevantes es que el modelo asume que los cambios que se van produciendo a lo largo del tiempo, según se analizan varios grupos de edad, son cambios esencialmente lineales, es decir, que el cambio producido entre las personas de 6 a 7 años es comparable a los cambios que se producen entre las personas de 8 respecto a las de 9 años. Desgraciadamente, asumir este principio es excesivo para muchos procesos psicológicos. Cambios producidos entre determinadas edades cronológicas no tienen por qué tener la misma relevancia o significado a lo largo de todas las edades mentales ni, por supuesto, al revés[3].

3. Baremación

La baremación, también conocida como escalamiento de una medida, es un elemento de gran relevancia en el proceso de elaboración y uso de un instrumento de medida. Como se ha visto en el presente capítulo, la transformación de una puntuación ayuda al usuario a inyectar significado a un determinado valor de acuerdo con el comportamiento de un conjunto de observaciones. Si el instrumento dispone de un manual, este debería reportar los baremos de la medida, con las transformaciones necesarias si es pertinente, además de las indicaciones básicas para interpretar debidamente cualquier puntuación. En otros casos no habrá manual de referencia, pero sí probablemente una o más publicaciones periódicas a donde podremos remitirnos y hallar la información necesaria.

Independientemente del tipo de publicación con la que el baremo de una medida se reporte, para que toda interpretación que se derive de este se haga se-

3. En Fancher (1987) encontraréis un texto interesante, en el que no solo se aborda el concepto de edad mental sino que también se hace un repaso histórico del concepto de inteligencia, así como de algunas controversias en torno a su medida.

gún un elevado estándar de calidad, es necesario que la muestra que proporciona los datos que permiten satisfacer una transformación (por ejemplo, medias y desviaciones típicas), y que recibirá el nombre de *muestra normativa*, sea una muestra relevante, representativa y homogénea.

Relevante porque hay que garantizar que la muestra que conforma los baremos de un instrumento está haciendo referencia a una población que es significativa de acuerdo con la satisfacción de unos objetivos específicos. Tal y como se ha visto durante la exposición de las puntuaciones estandarizadas, el sentido que puede tener una determinada puntuación tipificada debe ser entendido siempre dentro del marco de la población que proporciona los parámetros de referencia. No hay una población *a priori* ideal sobre la que desarrollar el baremo de un instrumento de medida. La población ideal depende del tipo de persona que se desea describir y, sobre todo, de los términos en los que se proponga hacerlo.

Probablemente todo el mundo estaría de acuerdo en que si se quiere inferir algo relevante sobre el grado de conocimiento de los jóvenes de un instituto sobre la Revolución Industrial sería conveniente poder disponer de una amplia muestra de otros jóvenes provenientes de otros institutos con los que contrastar el rendimiento. No obstante, no siempre es posible identificar una única población respecto a la que inferir satisfactoriamente el estado de un individuo.

De acuerdo con la misma línea argumental expuesta al hablar del neuroticismo y TDAH, si se desea interpretar la intensidad de la sintomatología depresiva de una persona mediante un determinado instrumento clínico, una puntuación concreta no tiene la misma consideración si los baremos empleados provienen de la población en general, de varios centros de atención primaria o de unidades de hospitalización psiquiátrica. *A priori*, las tres poblaciones son igualmente válidas, en cuanto que en todas ellas puede ser pertinente obtener datos normativos e interpretar la gravedad de la sintomatología depresiva de esta persona. Ahora bien, no es lo mismo ubicarse en un percentil de 70 tomando como muestra normativa población general que población clínica. La excepcionalidad (o no) de la sintomatología de este individuo dependerá necesariamente de cuál sea la población de referencia expresada mediante los baremos, y esto condicionará la interpretación que se pueda derivar, así como las decisiones que habría que tomar al respecto (McIntire y Miller, 2007). Hay que delimitar debidamente, a modo de objetivos de trabajo claros, qué debe ser evaluado y, sobre todo, con qué finalidad.

Se ha afirmado también que la muestra constitutiva de los baremos ha de ser *representativa*. Una vez definida cuál debe ser la población diana, la estrategia empleada para seleccionar a las personas de la población que formarán parte de la muestra normativa se transforma en el foco de atención principal. De acuerdo con los estándares que proporciona la American Psychological Association (APA, 1999), hay que exponer con cuidado y claridad cuál ha sido la población objetivo y, a la vez, la muestra que se ha seleccionado. Solo así es posible delimitar el tipo de inferencia a la que se puede llegar cuando unos determinados baremos son empleados.

Un muestreo probabilístico es la mejor estrategia metodológica disponible para asegurar la obtención de una muestra representativa, dado que solo es en este contexto aleatorio donde cualquier individuo de la población dispone de idéntica probabilidad de formar parte de la muestra. Y tal y como expone Aiken (1996), el muestreo aleatorio estratificado debería ser la estrategia de elección. Este tipo de muestreo debería proporcionar la debida representatividad para todos los parámetros que han sido considerados relevantes. En este sentido, la edad y el género suelen ser algunos de los factores que más frecuentemente articulan un baremo, pero cualquier otra variable puede ser objeto de interés y por lo tanto de estratificación. El ámbito de estudio y los objetivos que deben ser resueltos condicionarán el tipo de variables que se someten a escalamiento y el grado de estratificación con el que estas variables deberían expresarse.

Desgraciadamente no siempre se explicita en los textos especializados, pero no se puede dejar de mencionar, que no suele ser fácil poder satisfacer un muestreo estrictamente probabilístico. Son muchas las condiciones que, en contexto aplicado, reducen el éxito de un muestreo aleatorio. Considérese por ejemplo la presencia de personas institucionalizadas a las que difícilmente se tendrá acceso (prisioneros, personas hospitalizadas, etc.), aquellas que no tienen las aptitudes necesarias para dar respuesta a una determinada medida (analfabetas, con problemas de comprensión, etc.), las evaluadas en contextos en los que puede haber una alta deseabilidad social (selección de personal, presencia de preguntas comprometidas, etc.), personas que simplemente no quieren colaborar, etc. Los baremos de la mayoría de los instrumentos de evaluación, si no todos, están sujetos a circunstancias que limitan la generalización de sus valores. El esfuerzo de todo investigador debería dirigirse a minimizar o atenuar dentro de lo posible

estos sesgos y, en todo caso, informar siempre detenidamente a los usuarios del instrumento respecto a qué se ha hecho y cómo se ha resuelto.

Cuando se habla de la *homogeneidad* de la muestra se está haciendo referencia a la necesidad de controlar la presencia de efectos de confusión o extraños que puedan sesgar o incluso invalidar las interpretaciones que se deriven de esa muestra. Las personas que serán evaluadas mediante una determinada medida deberían ser comparables a las personas que formaron parte de la muestra de participantes que constituyeron el baremo de esta. Discrepancias consistentes entre las dos muestras pueden derivar en inferencias que sobreestimarían o infraestimarían el estado real de las personas evaluadas.

Estas mismas consideraciones justifican la necesidad de que haya, también, una continua actualización de los datos que estructuran los baremos. Se debería poder disponer de baremos que se sometan a revisiones con una determinada regularidad. Y es que habría que asumir por defecto que los datos contenidos en un baremo caducan. Las variables que son objeto de interés son dinámicas y, por lo tanto, cambiantes, moduladas por el decurso de los tiempos. Son sensibles a las transformaciones científicas, sociales y culturales que se van produciendo y, por esta razón, todo baremo debería poder reflejar fielmente el estado actual del concepto que es objeto de consideración. A pesar de que las consideraciones teóricas y empíricas son diferentes, emplear en la actualidad baremos de calidad de vida en muestras y bajo criterios de hace treinta años, por ejemplo, sería comparable a emplear baremos de altura incluyendo a personas evaluadas durante los años sesenta. En cualquiera de las dos opciones se estaría cometiendo un error o, en el mejor de los casos, un sesgo difícilmente asumible.

La adaptación del instrumento a las características lingüísticas y culturales de las personas que han de ser evaluadas debe ser también motivo de una necesaria atención. En la mayoría de los casos los instrumentos de evaluación que son empleados en psicología en nuestro entorno cultural tienen su origen en el mundo anglosajón. La ventaja que presenta esta circunstancia es que permite que, una vez traducido un instrumento, resulte más fácil la homologación del instrumento desde un punto de vista internacional, a la vez que facilita la satisfacción de estudios de tipo transcultural en los que un mismo rasgo contrasta entre varios grupos culturalmente diferentes. No obstante, hay que tener presente que traducir un instrumento es condición necesaria pero no siempre suficiente. Siguiendo las indicaciones de los estándares de la American Psychological Asso-

ciation (APA, 1999), es necesario que haya una adecuada adaptación cultural que sea sensible a las peculiaridades de los individuos que serán evaluados con el objetivo. Una correcta adaptación de la medida constituye un objetivo dirigido, de nuevo, a proporcionar estimaciones carentes de sesgo y que, por lo tanto, permitan inferir con las suficientes garantías de validez y fiabilidad el estado de los individuos.

En la configuración del baremo de un instrumento es posible emplear todas las estrategias expuestas en el presente capítulo. Los datos descriptivos, dispuestos en tablas, se pueden expresar de un modo distinto, por ejemplo mediante percentiles, y agrupados según estratificaciones por edad (Rodríguez et al., 2012). En otros casos, los percentiles se dispondrán sin estratificación (Alegret et al., 2012; Romero et al., 2000). En algunas publicaciones se ofrecerán medianas y desviaciones típicas que permitirán, si se deseara, una transformación en puntuaciones estandarizadas. Y lo pueden hacer según dos factores, por ejemplo, por edad y género (Gomà-i-Freixenet y Valero, 2008), o bien sin estratificación alguna (Aluja et al., 2008). Otros proporcionarán puntuaciones estandarizadas derivadas, a modo de puntuaciones T, y de manera estratificada (Yueh-Hsien et al., 2012). Las combinaciones son evidentemente múltiples, no son más que algunos de los muchos ejemplos que pueden ser presentados. Queda en manos del investigador proporcionar la información necesaria para justificar una u otra opción.

4. Equiparación de puntuaciones

Equiparar las puntuaciones de dos o más medidas consiste en establecer una correspondencia entre las puntuaciones de estas. Esta estrategia debe permitir que, recogida la puntuación en una de las medidas, sea posible estimar la puntuación de las otras medidas.

Uno de los contextos en el que puede ser especialmente interesante resolver equiparaciones entre tests es aquel en el que es necesario medir las aptitudes de

personas repetidamente y en el que, por lo tanto, se pueden dar procesos de aprendizaje o memoria, o simplemente cansancio, que podrían conducir hacia una más que probable estimación sesgada de las capacidades de las personas. En estas condiciones sería interesante poder disponer de dos o más formas alternativas de la medida que permitiera su uso indistintamente. No obstante, las condiciones que deberán ser satisfechas para una equiparación adecuada son las siguientes:

a) Que se mida la misma característica.

b) Que las dos medidas sean igualmente fiables.

c) Que la transformación sea invertible, es decir, la transformación debe ser posible de A a B, pero también de B a A.

Dos tipos de transformaciones son posibles: la equiparación horizontal y la vertical. El primer caso se daría cuando el rasgo de interés debe mantener su dificultad entre las diferentes medidas alternativas. Un solo grupo de sujetos suele ser el que, generalmente, proporciona las dos o más medidas. En el caso vertical, en cambio, se asume que la dificultad entre formas alternativas es diferente. En esta última circunstancia, son evaluados dos o más grupos de personas que presentan capacidades diferenciadas.

Los distintos diseños que permiten crear equiparaciones entre medidas son los siguientes: los de grupo único, los de grupos equivalentes, los tests de anclaje.

En el caso de los diseños de *grupo único*, un mismo grupo de personas extraído aleatoriamente de la población destinataria de las medidas equivalentes es evaluado mediante las diferentes medidas de interés. Como estrategia de control, para evitar procesos de aprendizaje y memoria, y por lo tanto un posible sesgo como consecuencia de haber administrado primero una forma alternativa y posteriormente la otra, se recomienda que las muestras respondan a las medidas de manera contrabalanceada. Así, se puede dividir la muestra aleatoriamente en dos grupos y administrar a cada subgrupo las distintas formas alternativas en un orden diferente, por ejemplo, en un grupo la secuencia A-B y en el otro la B-A.

En los diseños que implican *grupos equivalentes* son dos los grupos que se seleccionan aleatoriamente, pero en este caso es administrada una medida a la primera muestra y otra a la segunda. Poder asumir con garantías que las dos muestras de sujetos son comparables, por ejemplo, mediante una asignación

aleatoria de los sujetos, es condición necesaria para asegurar la validez del procedimiento.

La ventaja de emplear dos grupos de personas es que, en comparación con la estrategia anterior, reduce el efecto cansancio y no expone a los participantes a procesos de aprendizaje o memoria que podrían ser arrastrados entre las diversas medidas, dado que los participantes solo se exponen en una de ellas. La limitación más importante de esta aproximación es que, en general, exige incluir una cantidad de participantes superior a la necesaria en el caso del primer diseño.

Los *tests de anclaje* parten, también, de disponer de dos grupos de participantes que reciben cada uno de ellos una de las medidas dirigidas a ser equiparadas. Sin embargo, en este caso, a cada muestra se le administran algunos ítems que forman parte de la medida que es administrada en la otra muestra. Estos ítems comunes pasan a ser así ítems de anclaje. La estrategia es especialmente interesante cuando no es posible garantizar que los dos grupos de personas sean comparables, situación frecuente en un contexto aplicado en el que no ha sido posible generar los dos grupos de personas de manera aleatoria. Bajo esta circunstancia, los ítems de anclaje proporcionarán una estimación del grado en el que las dos muestras resultan comparables[4].

Independientemente de cuál de las tres estrategias haya sido empleada para satisfacer la equivalencia entre medidas, dos son los métodos más comunes que pueden ser ejecutados para resolver la equiparación entre medidas: equipercentil y transformación lineal.

La *equiparación equipercentil*, que suele ser la estrategia más habitual, consiste en considerar como equivalentes aquellas puntuaciones directas que se vinculan a un mismo percentil. Así pues, si las puntuaciones directas 10 y 14 obtenidas con la medida A y B, respectivamente, se asocian ambas a un percentil 30, se considerará que las dos puntuaciones son equivalentes. El mismo razonamiento sería aplicable al resto de los pares de puntuaciones directas que convergieran en un mismo percentil.

Las consideraciones que deben ser tenidas en cuenta a la hora de aplicar esta estrategia son las mismas que ya han sido expuestas al hablar de los percentiles. Se trata, pues, de una estrategia de fácil comunicación y ejecución.

4. En el texto de Muñiz (2003) encontraréis más información respecto a esta estrategia.

La *transformación lineal* consiste en equiparar puntuaciones estandarizadas. Concretamente consiste en equiparar aquellas puntuaciones directas que convergen en una misma puntuación típica. Expresado de otro modo, y de acuerdo con la fórmula que permite calcular una puntuación típica *z*, se puede decir que:

$$\frac{X-\overline{X}}{Sx}=\frac{Y-\overline{Y}}{Sy}$$

De donde si se aísla *Y*:

$$Y=\frac{Sy}{Sx}(X-\overline{X})+\overline{Y}$$

Ejemplo

Consideremos el ejemplo de dos muestras equivalentes de personas extraídas aleatoriamente de una determinada población (diseño de grupos equivalentes). La primera muestra obtiene, en una medida de aptitudes aritméticas (*X*), una media de 10 y una desviación típica de 4. La segunda muestra, en la segunda medida también de aptitudes aritméticas (*Y*), obtiene una media de 8 y una desviación típica de 2. Según esta información, una puntuación de 9 en la medida *A* ¿a qué puntuación directa correspondería en la medida *B*?

$$Y=\frac{4}{2}(9-10)+8=6$$

Obtener un 9 en la medida *A* es equivalente a obtener una puntuación de 6 en la medida *B*.

En el caso del diseño de un solo grupo en el que ha sido empleada una estrategia de contrabalanceo, hay que tener en consideración el orden en el que fueron administradas las dos medidas (orden *A-B* y orden *B-A*). Bajo esta circunstancia se calcula la desviación típica y media en cada uno de los grupos de balanceo. Así pues, empezando por la desviación típica:

$$Sx=\sqrt{S_{x1}^2+S_{x2}^2}$$

$$Sy=\sqrt{S_{y1}^2+S_{y2}^2}$$

Donde los estadísticos incluidos en los radicales son las varianzas correspondientes al primer y segundo órdenes en la administración de las dos medidas.

En el caso de las medias:

$$\overline{X} = \frac{\overline{X}_{x1} + \overline{X}_{x2}}{2}$$

$$\overline{Y} = \frac{\overline{X}_{y1} + \overline{X}_{y2}}{2}$$

Siendo los estadísticos de los numeradores las medias correspondientes al primer y segundo orden de las dos medidas. Calculadas las medias y desviaciones típicas, solo falta aplicar la fórmula ya conocida que permite establecer la correspondencia entre las puntuaciones directas equivalentes de las dos medidas.

Ejemplo

Las medidas *X* e *Y* han sido administradas a dos grupos de personas de manera contrabalanceada. La medida *X*, cuando fue administrada en primer lugar, presentó una media de 10 y una desviación típica de 2. Cuando fue administrada en segundo lugar, la media fue de 8 y la desviación típica de 1. En el caso de la medida *Y*, administrada en primer lugar, presentó una media de 12 y una desviación típica de 3, mientras que administrada en un segundo orden su media fue 14 y su desviación típica 4. ¿Qué puntuación se correspondería en la medida *Y*, si en la medida *X* se obtuviera una puntuación de 7?

Calculando las desviaciones típicas de las medidas *X* e *Y* según el orden en el que fueron administradas, respectivamente:

$$2,24 = \sqrt{4+1}$$

$$5 = \sqrt{9+16}$$

En el caso de las medias, respectivamente:

$$9 = \frac{10+8}{2}$$

$$13 = \frac{12+14}{2}$$

Estableciendo la correspondencia entre las dos medidas, se llegaría a la conclusión de que la puntuación en la medida *Y* debería ser 9:

$$\frac{5}{2,24}(7-9)+13 = 8,54 \approx 9$$

La aplicación de esta forma de equiparación entre medidas debe ajustarse a las mismas consideraciones que fueron recogidas al hablar de las puntuaciones estandarizadas. Esto implica, por lo tanto, depositar el interés sobre la tendencia de respuesta de los individuos en las dos medidas y asumir que las dos distribuciones de valores objeto de equiparación se ajustan básicamente a una distribución normal.

Resumen

En la primera parte del presente capítulo se ha hablado de los percentiles, las puntuaciones estandarizadas y las normas cronológicas, puntuaciones transformadas comúnmente empleadas, dirigidas a otorgar interpretabilidad a las puntuaciones obtenidas en una medida. Los percentiles, calculados de acuerdo con la posición de una persona en un continuo y expresados como porcentajes, se convierten en una aproximación fácil de entender y por lo tanto de difundir, incluso a personas poco acostumbradas al uso de este tipo de estrategias matemáticas. Además, el hecho de ser esencialmente insensibles a los grupos de valores no ajustados a una distribución normal permite que se puedan considerar una transformación de amplio espectro, una estrategia útil y aplicable para la mayoría de las variables cuantitativas que son de interés en psicología.

Las puntuaciones estandarizadas presentan más limitaciones en términos de aplicabilidad. Requieren que las observaciones que son objeto de análisis se ajusten básicamente a una distribución normal. Y tal y como se ha recogido, son muchas las variables psicológicas que, evaluadas en una muestra concreta de personas, acaban sin satisfacer este modo de distribución. No obstante, cuando esta circunstancia puede ser asumida, la estandarización supone una aproximación diferenciada a la de los percentiles, que puede resultar especialmente inte-

resante. A diferencia de los percentiles, se focalizan en la tendencia de respuesta, la media de comportamiento que proporciona el grupo normativo. La desviación típica, la diferencia entre los individuos del grupo normativo, se convierte en la manera de cuantificar la discrepancia entre la puntuación de una persona concreta y ese comportamiento promedio.

Concluyendo el conjunto de estrategias de transformación, se ha expuesto la puntuación a modo de normas cronológicas o edad. En este caso, y de acuerdo con la tendencia de respuestas observadas de nuevo en un grupo normativo, constituido por edades diferentes, la puntuación de un individuo pasa a ser transformada e interpretada de acuerdo con esta variable cronológica. Su razonamiento, como en el caso de los percentiles, es también simple, fácilmente comunicable. No obstante, y tal y como se ha hecho notar, presenta inconvenientes que no se pueden menospreciar. Pese a estas limitaciones, ha sido una estrategia muy empleada en el ámbito de la medida de la inteligencia, lo que ha justificado el espacio dedicado a ella en el presente capítulo.

El baremo ha sido uno de los conceptos fundamentales del capítulo. Gracias a él el usuario de un instrumento puede realizar inferencias respecto a la puntuación de un individuo, de acuerdo con el perfil de comportamiento de una muestra normativa, y empleando tablas que establecen esta correspondencia, muchas veces mediante el uso de puntuaciones transformadas. No obstante, más allá de las consideraciones que se derivan del uso de una u otra estrategia de transformación, el elemento más relevante que hay que tener en consideración es la necesaria y adecuada justificación de cuál debe ser esta muestra normativa que estructura este baremo. Esta muestra determinará fuertemente las inferencias que se puedan derivar de ella y podrá condicionar, cuando se esté trabajando en un contexto aplicado, el sentido de las decisiones que deban ser tomadas.

El capítulo concluye con algunas de las estrategias más frecuentes empleadas dirigidas a equiparar puntuaciones entre medidas diferentes que miden un mismo objeto psicológico. La equiparación equipercentil o la transformación lineal devienen estrategias que pueden ser especialmente útiles, por ejemplo, en aquellos contextos de trabajo en los que una evaluación repetida de una misma variable en un mismo individuo puede conducir a una estimación sesgada. Las ventajas y limitaciones que el usuario debe tener presentes cuando se desee emplear alguna de estas dos estrategias son homólogas a las descritas en el caso de los percentiles o de las puntuaciones estandarizadas.

Capítulo V

Análisis de los ítems

Albert Bonillo

El objetivo de este capítulo es introducir al lector en el tema del análisis de ítems. Tradicionalmente, el estudio de la psicometría se ha focalizado más hacia las propiedades de los instrumentos de medida, que preguntan sobre aspectos opinativos y constructos psicológicos, que hacia los instrumentos que miden conocimiento o habilidad. Sin embargo, el psicólogo de a pie trabaja tanto o más con los segundos que con los primeros.

Queremos que el lector sepa, desde el principio, que este capítulo no lo hará experto en ninguno de los aspectos que en él se tratan. Le proporcionará, deseamos y esperamos, una buena introducción a cada uno de los temas, pero es fácil que por placer –o necesidad profesional, o ambas– necesite profundizar en algunos de los aspectos tratados. Recomendaremos textos que sí los traten en profundidad.

El capítulo se inicia precisamente distinguiendo entre instrumentos en función de su objetivo. En segundo lugar, y ya centrados en pruebas de ejecución máxima, mostraremos cuáles son los aspectos que hay que tener en cuenta en la construcción de sus ítems. En el tercer apartado veremos cómo analizar las propiedades psicométricas de la prueba y de los ítems a partir de la teoría clásica de test (TCT). Veremos los conceptos de dificultad y discriminación, y aprenderemos a valorar si un ítem es correcto o quizá necesita una revisión. En el cuarto apartado veremos una introducción a la teoría de respuesta al ítem (TRI), que es una alternativa de análisis a la TCT. Veremos la TRI de manera más sucinta que la TCT. El modelo de TRI resuelve problemas teóricos de la TCT, pero los cálculos de esta son más sencillos y fácilmente aplicables que los de aquella.

Dejaremos para el final las conclusiones que resuman todo lo presentado.

1. Tipos de pruebas

Es tradicional que, cuando desde el ámbito de la psicología hablamos de una prueba –o de un instrumento de medida– pensemos de inmediato en una encuesta de opinión, un test de personalidad o similar. Desde el punto de vista del tipo de prueba, estas que hemos mencionado no son distintas del cuestionario de satisfacción sobre el servicio que encontramos a la salida de muchos hoteles. Pretenden medir, en una persona, el valor determinado de un constructo cuya existencia se presupone.

Caso distinto es una prueba que pretenda ordenar a los mejores candidatos a un puesto de trabajo. En este contexto, donde es de suponer que existe un criterio –ser un buen trabajador para el puesto ofertado–, la medida del constructo puede pasar a un segundo plano. El objetivo del instrumento es que cada uno de los ítems optimice la correcta clasificación de las personas. Veamos, pues, qué características tienen las pruebas en función de lo que pretenden.

1.1. Pruebas de ejecución típica frente a pruebas de ejecución máxima

Si clasificamos las pruebas por su objetivo, distinguiremos entre dos tipos básicos. Denominamos pruebas de ejecución típica –o de ejecución de rasgos– a aquellas que miden aspectos no escalables, o dicho de otra manera, a aquellas cuyas preguntas no tienen respuestas correctas ni erróneas, sino que se trata de aspectos de opinión, de preferencia o similar. Por el contrario, llamamos pruebas de ejecución máxima a aquellas que evalúan constructos que sí son escalables, y que son aquellos en los que tiene sentido hablar de respuestas correctas y erróneas. Un examen, un test de inteligencia o cualquier instrumento que mida aptitud sería clasificado dentro de este epígrafe.

Aunque todos los conceptos que hemos visto hasta ahora en capítulos anteriores –fiabilidad, validez y transformación de las puntuaciones obtenidas– son aplicables a ambos tipos de instrumentos, las estrategias para su estudio suelen variar ligeramente y se suelen estudiar aplicándolos a las pruebas de ejecución típica. Es cierto que, por ejemplo, un test de inteligencia debe ser fiable, pero

puede no tener demasiado sentido administrarlo dos veces en unas pocas semanas, ya que los participantes podrían haber obtenido la respuesta correcta en el tiempo transcurrido y contaminar así los resultados. Sin embargo, sí tiene sentido repetir un test de personalidad con pocos días de diferencia y comprobar de ese modo si la medida del instrumento es tan estable como se supone que es el constructo medido. En definitiva, las características que se deben estudiar dependen, cómo no, del objetivo del instrumento.

En muchas ocasiones, el psicólogo profesional no utiliza instrumentos estandarizados, sino que debe crear él mismo el instrumento. Si el lector trabajara en el departamento de recursos humanos de una multinacional y esta le pidiera una prueba para ocupar un puesto muy específico, ¿qué haría? Tras comprobar que esta prueba no existe en el mercado debería crearla. Y debería hacerlo teniendo en cuenta qué se pretende hacer con esa prueba: seleccionar al mejor trabajador para ese puesto. ¿Y a partir de ahí? Supongamos que ese puesto requiere ciertos conocimientos. El psicólogo debería construir una prueba que, a partir del número mínimo de ítems, pueda seleccionar al mejor de los candidatos.

Aprendamos, pues, qué debe tenerse en cuenta cuando (no) hay (más remedio) que crear una prueba.

2. Directivas en la construcción de ítems

El lector ya sabe que el objetivo principal de este capítulo es mostrar cómo medir la calidad de un test de rendimiento. Ahora bien, no debemos eludir explicar qué hay que hacer para construir correctamente una prueba. Creemos que el trabajo de un psicólogo no debe ser únicamente valorar si la prueba está mejor o peor hecha, sino que también tiene mucho que aportar en su construcción. Cambiando totalmente de ámbito: ¿no sería extraño que un arquitecto valorara edificios si no aprendiera primero a construirlos?

Existen varios trabajos que exponen de manera muy exhaustiva cuáles son las directivas que hay que seguir para construir correctamente una prueba de ejecución máxima. Uno de los primeros y más conocidos es el de Haladyna, Downing y Rodríguez (2002). Se trata, ni más ni menos, que de 31 criterios que

seguir, clasificados por apartados. Estos criterios se refieren al contenido de la pregunta (por ejemplo, cada ítem debe medir un único conocimiento), al formato, al estilo (recomienda ítems cortos), al enunciado (tienen que evitar las negaciones) y a las opciones de respuesta (recomienda evitar la opción "Todas las anteriores son correctas/incorrectas").

Personalmente, preferimos los criterios de Moreno, Martínez y Muñiz (2004). Son menos (doce), son mucho más claros y más fáciles de aplicar. Como podéis ver en la tabla 1, ahora los aspectos que hay que valorar son tres: elección del contenido, su expresión y opciones de respuesta.

Tabla 1. Nuevas directrices para la construcción de ítems de elección múltiple

A. Elección del contenido que se desea evaluar
1. Debe ser una muestra representativa del contenido recogido en una tabla de especificación, evitando ítems triviales.
2. La representatividad deberá marcar lo sencillo o complejo, concreto o abstracto, memorístico o de razonamiento que deba ser el ítem, así como el modo de expresarlo.
B. Expresión del contenido en el ítem
3. Lo central debe expresarse en el enunciado. Cada opción es un complemento que debe concordar gramaticalmente con el enunciado.
4. La sintaxis o estructura gramatical debe ser correcta. Evitar ítems demasiado escuetos o profusos, ambiguos o confusos, cuidando además las expresiones negativas.
5. La semántica debe ajustarse al contenido y a las personas evaluadas.
C. Construcción de las opciones
6. La opción correcta debe ser solo una, acompañada por distractoras plausibles.
7. La opción correcta debe estar repartida entre las distintas ubicaciones.
8. Las opciones deben ser preferiblemente tres.
9. Las opciones deben presentarse usualmente en vertical.
10. El conjunto de opciones de cada ítem debe aparecer estructurado.
11. Las opciones deben ser autónomas entre sí, sin solaparse ni referirse unas a otras. Por ello, deben evitarse las opciones "Todas las anteriores" y "Ninguna de las anteriores".
12. Ninguna opción debe destacar del resto ni en contenido ni en apariencia.

Fuente: Tomado de Moreno, Martínez y Muñiz (2004)

En el contenido, deben preguntarse cosas fundamentales. Parece obvio, pero ¿cuántos exámenes recordamos en los que se nos preguntaban algunas cuestiones que aparecieron poco (o nada) en clase? Una prueba debería contener solo

(pero todos) los conceptos fundamentales de la materia que valora. La creencia de que al preguntar cuestiones menores, en el fondo, estamos obligando al alumno a estudiar toda la materia es absurda y favorece el azar. Respecto al azar, recordad a aquellos alumnos que solo estudiaban medio programa –o menos– y confiaban en tener suerte el día del examen.

Sobre la expresión, las tres cuestiones apuntadas son obvias, pero de nuevo no siempre se cumplen.

Un ejemplo paradigmático de Moreno, Martínez y Muñiz (2004) muestra que es mejor redactar este ítem:

En física, se denomina sublimación a un cambio de materia:

1. Sólida a gaseosa.
2. Líquida a sólida.
3. Gaseosa a líquida.

que este:

En física, sublimación:

1. Supone un cambio de materia sólida a materia gaseosa.
2. Se refiere a un cambio de materia líquida a materia sólida.
3. Consiste en un cambio de materia gaseosa a materia líquida.

Sobre las opciones de respuesta, destacaremos la recomendación de que las opciones sean independientes entre sí, lo que automáticamente conlleva no usar los célebres "Todas/Ninguna de las anteriores". Es obvio que para rechazar una opción como "Todas las anteriores son correctas" solo necesitamos saber que una de las otras opciones no lo es. Así, de un plumazo, podemos eliminar dos opciones de las posibilidades y la elección se facilita mucho. Si el test tiene tres opciones, ya conocemos la respuesta, y si tiene cuatro, incluso podemos arriesgarnos a contestar al azar entre las dos restantes[1].

1. Si deseáis profundizar en este tema, os recomendamos acudir al texto original de Moreno, Martínez y Muñiz (2004), en el que, en un tono muy didáctico y con ejemplos muy accesibles, encontraréis una explicación muy exhaustiva de cada uno de los criterios.

Ejemplo de prueba de ejecución máxima

A partir de las directrices mostradas, y para ilustrar con un ejemplo concreto y cercano los conceptos que se presentarán en este capítulo, hemos construido el siguiente examen. Contiene diez preguntas sobre este mismo capítulo y la opción correcta está resaltada en negrita[2].

1. La dificultad (ID) es un índice que indica la probabilidad de…

A. acertarlo.
B. fallarlo.
C. contestarlo.

2. El valor de discriminación de un ítem (ID) debe ser…

A. negativo.
B. distinto de 0.
C. positivo.

3. Un distractor debería tener discriminación…

A. positiva.
B. negativa.
C. cercana a 0.

4. Un test de personalidad es una prueba de...

A. ejecución máxima.
B. ejecución típica.
C. rendimiento.

5. La fórmula para calcular *IDc* es...

A. $\frac{A-\frac{E}{1-K}}{N}$

2. En el último apartado de este capítulo se detalla un ejemplo de respuestas (ficticias) de un grupo de veinte alumnos a esta prueba, junto a los cálculos de la mayoría de los índices a los que haremos referencia en este texto.

B. $\frac{A-\frac{K}{E-1}}{N}$

C. $\frac{A-\frac{E}{K-1}}{N}$

6. El modelo de TRI calcula, a partir del conocimiento,...

A. la puntuación total esperada.
B. la probabilidad de acertar un ítem.
C. la discriminación del test.

7. Los parámetros *a*, *b* y *c* de la TRI indican, respectivamente,…

A. discriminación, **dificultad**, **pseudoadivinación**.
B. dificultad, discriminación, pseudoadivinación.
C. pseudoadivinación, discriminación, dificultad.

8. Si, por nivel de dificultad, solo pudiéramos tener ítems de un tipo, estos deberían ser, generalmente,…

A. fáciles.
B. difíciles.
C. medios.

9. Un ítem que pregunte sobre un aspecto del temario difícil debería ser...

A. fácil.
B. medio.
C. difícil.

10. La evaluación del sesgo pretende…

A. hacer más justas las pruebas.
B. evaluar la dificultad de los ítems.
C. aumentar la fiabilidad de la prueba.

3. Teoría clásica

Existen dos grandes modos de acercarse al análisis de ítems. Distinguiremos, pues, entre la teoría clásica de test (TCT) y la teoría de respuesta al ítem (TRI). La primera la estudiaremos en este apartado y la segunda, en el siguiente. ¿Qué supuestos tiene la TCT? Aunque se estudia en profundidad en el apartado de fiabilidad, se resumen en la ecuación

$$X=\text{V}+\text{E}$$

Esta implica que la puntuación que una persona obtiene al contestar un instrumento de medida (X) contiene el denominado "nivel verdadero" de esa persona (V) y una parte de error. El objetivo de la TCT es, pues, medir y minimizar ese error, lo que implica analizar la fiabilidad de la medida. Como no podía ser de otra manera, y siempre bajo estos supuestos, todos los indicadores de calidad de los ítems dependen de la muestra de personas que los han contestado.

Veamos, ahora, las principales propiedades que se han de medir de un ítem.

3.1. Dificultad

El *índice de dificultad* de un ítem (ID) es la proporción de personas que lo contestan correctamente. Es decir,

$$ID = \frac{A}{N}$$

A: Número de personas que aciertan el ítem.

N: Número total de personas que lo contestan.

Al tratarse de una proporción, ya que los acertantes son un subconjunto de los que contestan, es obvio que sus valores fluctúan entre 0 y 1, y frecuentemen-

te se expresan como un porcentaje. Paradójicamente, los valores cercanos a 1 indican una baja dificultad –debería llamarse pues índice de facilidad y no de dificultad– y valores cercanos a 0 indican dificultad máxima. Podemos ver en la fila titulada como ID de la hoja de cálculo anexa las dificultades de los ítems del ejemplo de prueba de ejecución máxima.

La fórmula anterior presenta un problema: no tiene en cuenta que una parte de los aciertos se dan por puro azar. Al tratarse de preguntas con alternativas cerradas, es lógico pensar que una parte de los acertantes no conocían la respuesta y que si han acertado es "solo" porque han elegido una de las alternativas, no porque "sepan" la respuesta. El problema que se nos plantea es desconocer cuántos lo han hecho. La solución es muy intuitiva. Si cien personas que no saben japonés contestaran un examen redactado en esa lengua con preguntas de cuatro alternativas, ¿cuántas preguntas esperaríamos que acertaran? La esperanza matemática –y el sentido común– nos dice que veinticinco. ¿Y si hubiera cinco alternativas de respuesta? Obviamente, veinte. Por lo tanto, el número de aciertos total –y por ende, la probabilidad de acertar una pregunta– depende en cierta medida del número de alternativas de respuesta.

Por tanto, es recomendable utilizar la siguiente fórmula.

$$IDc = \frac{A - \frac{E}{K-1}}{N}$$

A: Número de personas que aciertan el ítem
E: Número de personas que fallan el ítem
K: Número de alternativas (u opciones) de respuesta
N: Número total de personas que lo contestan

Inmediatamente, observamos que la diferencia entre ambas fórmulas es que en la segunda se resta de *A* un número que se obtiene de dividir los errores (*E*) entre el número de alternativas erróneas (*K*-1).

Para aprehender este concepto, observemos los resultados del ítem 4. Lo han acertado 8 de los 20 participantes. Su dificultad sin corregir es por tanto del 40%. Ahora bien, al corregir el índice por el acierto al azar [(12/[(3-1)):20]=3], obtenemos un 30%, esto es, podemos suponer que el 30% de las personas lo han

acertado de casualidad. Vale la pena observar que con el primer cálculo obtendríamos una dificultad del 40% (podríamos etiquetarlo como de dificultad media), mientras que con el segundo obtendríamos una *IDc* = 10% (alta dificultad).

Aunque conceptualmente no tiene sentido, puesto que sigue siendo una proporción, el *IDc* puede ser inferior a 0: en ese caso, se asigna *IDc* = 0. Esto es lo que ocurre con el ítem 10, que tiene una dificultad corregida de –5%, lo que no tiene sentido. Observemos también cómo los ítems 1 y 3 son perfectamente inútiles, el primero por fácil y el segundo por difícil.

Un ítem que todos aciertan o que todos fallan no sirve para nada más que para perder el tiempo contestándolo. Si todo aquel que responde acierta, es como si regaláramos a todos los alumnos una parte de la puntuación. Y si todos lo fallan, es como si los penalizáramos. Supongamos una prueba que tiene 10 ítems y una puntuación teórica entre 0 y 10. En el primer caso que hemos expuesto, la puntuación real podría fluctuar entre 1 y 10, y en segundo entre 0 y 9. Está claro que esto no habla bien de las propiedades de la prueba.

Una vez que sabemos la dificultad de un ítem, planteémonos, ¿cómo deberían ser las dificultades de todos los ítems de una prueba? Como dice la directriz dos de Moreno, Martínez y Muñiz (2004), la dificultad de un ítem debe relacionarse con la del concepto que recoge. Esto es, si un contenido es fácil, el ítem debe ser fácil. Por tanto, una prueba que mide contenidos diversos debería tener ítems de todas las dificultades, y éstas deberían corresponderse a la dificultad de los conceptos medidos.

Una propuesta, realizada por nosotros (Bonillo, 2012) es mostrar en un gráfico la dificultad (eje Y) de los ítems de una prueba (ordenados de menor a mayor dificultad en el eje X) y observar si la pendiente es cercana a los 45º. Esto es, la línea que une los puntos de estas dificultades debería cruzar en diagonal el gráfico.

La siguiente figura muestra esta idea aplicada a ítems de los exámenes de acceso a la formación sanitaria especializada (las célebres pruebas de médico, farmacéutico y enfermera interno residente, MIR, FIR y EIR, respectivamente, para dos años, el 2005 y el 2006). Si sumamos ambas convocatorias, se presentaron a ellas más de 23.000 aspirantes y sus resultados son muy relevantes, ya que los aspirantes que las superan serán los futuros médicos, farmacéuticos y enfermeras especialistas.

Figura 1. Ítems y dificultad de las pruebas FSE

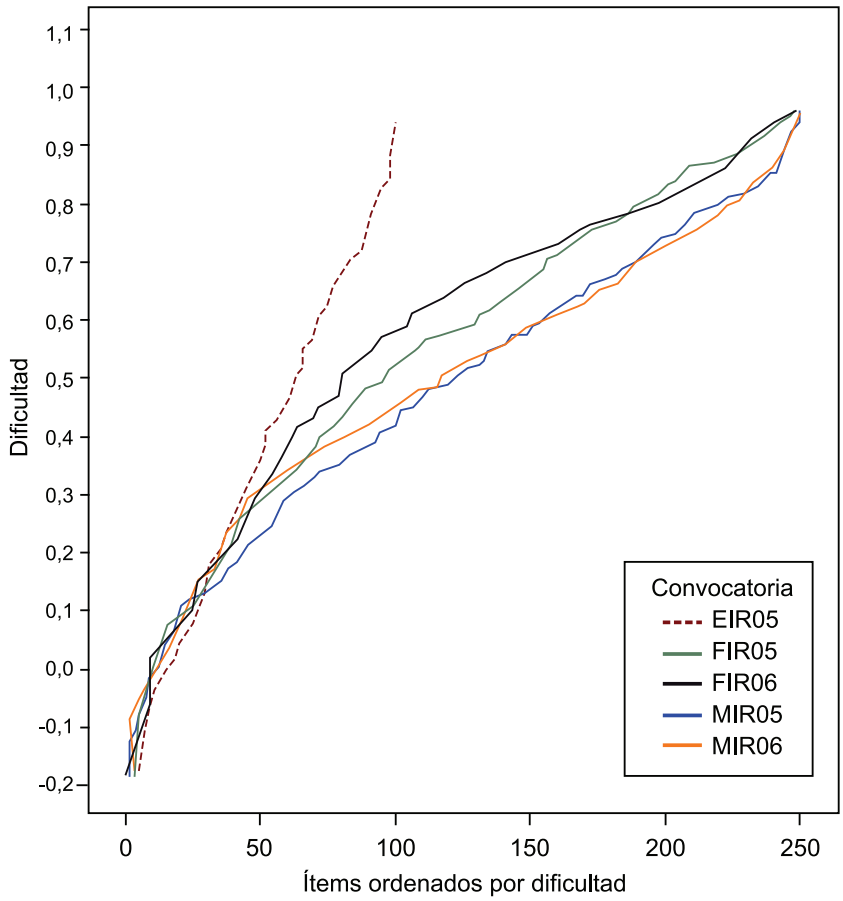


Figura 1. Ítems y dificultad de las pruebas FSE

Podemos ver en el gráfico varias cuestiones que merecen la pena destacarse. En primer lugar, observamos que hay dificultades negativas, y que estas se explican por aplicar la corrección del azar a ítems muy difíciles. En segundo lugar, cabe tener en cuenta que la prueba de enfermería cuenta con 100 ítems (frente a los 250 del resto de las convocatorias), lo que explica la evidente diferencia en las pendientes. En tercer lugar, las curvas de las pruebas de farmacia están, en el gráfico, "más altas" que las de medicina, lo que indica que son pruebas más fáciles. En cuarto lugar, las curvas son muy semejantes intraprogramas, es decir, la dificultad de las pruebas es muy semejante año tras año[3].

3.2. Discriminación

¿Es suficiente saber si un ítem es fácil o difícil para decidir si es adecuado o no? Intuitivamente, podríamos pensar que sí, pero estaríamos equivocados. De hecho, si tuviéramos que destacar una propiedad psicométrica de los ítems sobre el resto, esta sería la discriminación. Si un ítem no discrimina, no es útil para la medición, y ese es el objetivo para el que fue redactado.

Como su nombre indica, entendemos como discriminación la capacidad de un ítem de distinguir entre las personas que tienen un buen rendimiento en el test, respecto a las que lo tienen malo.

¿Quiénes deben contestar correctamente una pregunta de examen? No es tan importante si son muchos o pocos alumnos como que los acertantes sean, en general, "de los buenos alumnos". ¿A qué nos referimos cuando decimos "los buenos"? A aquellos que tienen una alta puntuación en la prueba. Es decir, un ítem debe ser más acertado entre aquellos que han obtenido una alta puntuación en la prueba que entre los que no la tienen. Obviamente, una pregunta no puede ser buena si solo la aciertan los peores alumnos: debe ocurrir lo contrario.

El índice de discriminación más popular es el índice D, conocido también como índice basado en las proporciones de aciertos.

$$D = Pa - Pb$$

P_a: Proporción de personas del grupo de alto rendimiento que acierta el ítem
P_b: Proporción de personas del grupo de bajo rendimiento que acierta el ítem

Las proporciones se calcularían como hemos visto en la primera fórmula presentada y, de nuevo, puede expresarse como porcentajes. Pero ¿quiénes son los

3. Si deseáis conocer con más detalle la propuesta, que aquí solo apuntamos, podéis consultar el artículo original (Bonillo, 2012).

alumnos de alto y bajo rendimiento? Existen varias maneras de definir el punto de corte de la puntuación total en la prueba para hacer esta clasificación. Por un lado, es frecuente utilizar la media de la puntuación total en la prueba, lo que crea dos grupos de igual tamaño. Esta estrategia tiene como ventaja que todos los participantes participan en el cálculo, pero tiene como claro inconveniente que los grupos son poco extremos. Intuitivamente comprobamos que dos personas con rendimiento muy semejante pueden estar en grupos diferentes solo por una pequeña diferencia.

Es preferible utilizar grupos más extremos para poder estudiar correctamente este índice. Kelley (1939) recomienda utilizar los percentiles superior e inferior del 27%. ¿Por qué 27% y no 25%? Aunque el artículo original demuestra que el 27% es ligeramente mejor que el 25%, en el ejemplo con respuestas ficticias que mostramos se utiliza el 25% como criterio para separar el grupo de rendimiento alto –se interpretaría como aquel que obtiene puntuaciones superiores al 75% de sus homólogos– del bajo –que reúne el 25% de las puntuaciones más bajas. Calcular el percentil 27 no siempre es sencillo, mientras que el 25 sí lo es, y las variaciones entre uno y otro son muy menores.

¿Cuáles son los límites de D? Es obvio que, teóricamente, puede fluctuar entre 1 y –1. El primer valor se daría solo cuando todas las personas del grupo superior acertaran y todas las del inferior fallaran. En valor –1 solo podría darse en el caso contrario, y entonces deberíamos sospechar si la respuesta considerada como correcta lo es. Ninguna de estas dos situaciones suele darse en la realidad.

¿Cómo debemos pues interpretar este índice? En primer lugar, solo valores positivos indican discriminación. Está claro que un ítem debe ser más acertado entre los mejores. Pero ¿qué valores indican una buena discriminación? Ebel (1965) propuso la siguiente clasificación, que debe ser tomada como orientación:

Tabla 2. Puntos de corte de los valores (en %) de discriminación (D) y su interpretación

D	Interpretación de la discriminación
> 40	Alta discriminación
30-40	Aceptable
20-30	Baja: se sugiere revisar el ítem
0-20	Mala: se elimina el ítem o se reforma profundamente
< 20	Inaceptable: eliminar el ítem

Un motivo para tomar la tabla anterior con precaución es que el índice *D* depende –y mucho– de la dificultad. Si un ítem es muy difícil, tendrá pocos acertantes (por definición), incluso en el grupo de alto rendimiento. Si P_a es baja, la *D* solo puede ser baja. No parece justo comparar *D* de ítems de dificultades muy diferentes. Una alternativa propuesta al índice *D* es calcular la diferencia de proporciones relativa, en lugar de la absoluta. Es decir,

$$Dr = (Pa - Pb) / P$$

Calculemos la discriminación de los ítem 4 y 5, que tienen la misma dificultad.

Ítem 4: $D = \left(\frac{4}{5} - \frac{0}{0}\right) \cdot 100 = 80\%$ y $Dr = \left(\frac{0{,}8 - 0}{0{,}8}\right) \cdot 100 = 100\%$

Ítem 5: $D = \left(\frac{2}{5} - \frac{0}{0}\right) \cdot 100 = 40\%$ y $Dr = \left(\frac{0{,}4 - 0}{0{,}4}\right) \cdot 100 = 100\%$

Para el ítem 4: el primer valor (*D* = 80%) debería interpretarse de la siguiente manera: los mejores aciertan el ítem un 80% más que los peores. O interpretarlo como una diferencia de probabilidades: es un 80% más probable acertar el ítem cuando se tiene un alto rendimiento (que si se tiene bajo). El segundo valor (*Dr* = 100%) se interpretaría como que los buenos aciertan un 100% más (respecto a los malos). En este caso, ambos valores hablan bien de la capacidad discriminativa del ítem. Para el ítem 5, los datos son parecidos (*D* = 40% y *Dr* = 100%). Así pues, ambos ítems serían más que aceptables.

Supongamos ahora un ítem muy difícil, que solo sea acertado por 1 de cada 20 participantes, y supongamos que este pertenece al grupo de alto rendimiento. Está claro que *D* = 1/6-0/6 = 0,17 = 17%. Según los criterios de Ebel tendríamos que eliminarlo, pero *Dr* = 100%, y por tanto tiene discriminación relativa máxima. ¿Qué deberíamos entonces hacer? No existe una respuesta absoluta a esta pregunta.

Si, por las características de la prueba, es aceptable tener un ítem tan difícil, este debería mantenerse ya que discrimina tanto como puede discriminar. Si, por el contrario, no es conveniente que tan pocas personas lo acierten, debería reformarse, pero la decisión –en este caso– depende por completo de la dificultad y no de la discriminación.

En resumen, hay que tener claro que la discriminación depende de la dificultad y no debe interpretarse *per se*. En términos estadísticos, la discriminación depende de la variancia de la dificultad.

Existen otros índices alternativos a los presentados para medir la discriminación. Uno de los más utilizados es la correlación ítem-test. Habitualmente, se utiliza el índice de correlación biserial-puntual, ya que permite cuantificar la relación entre una variable binaria –el acertar o no el ítem– y una variable de escala –la puntuación total de la persona en la prueba, idealmente sin tener en cuenta el ítem analizado. La fórmula y la lógica de esta prueba se encuentran ampliamente explicadas tanto en Muñiz (2003, p. 220) como en Martínez Arias (1992, p. 556), pero es muy sencillo ver que una alta correlación –cercana a 1– indica una gran discriminación del ítem, que valores cercanos a –1 indican lo contrario (donde los buenos fallan el ítem y los malos lo aciertan) y que valores cercanos a 0 indican que nada tiene que ver acertar este ítem con el conocimiento que mide el conjunto de la prueba.

3.3. Discriminación de los distractores

Un aspecto clave para el buen funcionamiento de un ítem es que sus distractores[4] realmente lo sean. Una alternativa que no es elegida por nadie –o casi– no está confundiendo a las personas que responden, y por tanto no es útil. Y a la inversa: una alternativa incorrecta que es muy elegida por los mejores quizá no es tan incorrecta como pensó el que la redactó.

¿Cómo se estudia el comportamiento de los distractores? El índice habitual vuelve a ser el índice *D* que ya hemos visto, pero en lugar de calcularlo a partir de quienes aciertan y fallan, se hace con quienes eligen cada una de las alternativas de respuestas.

Observad atentamente los resultados de administrar el ítem 5. La discriminación de la opción de respuesta B es más alta que la de la respuesta correcta, la C. Debemos interpretarlo como que los mejores eligen en mayor medida un distractor –como es la C– que la respuesta correcta –como es la B. Esto implica

4. *Distractores* es el nombre que se le da a las alternativas de respuesta incorrectas.

que debemos comprobar si la pauta pudiese contener un error. En nuestro caso no es así, y podemos atribuir al azar que la opción C haya sido tan elegida. Ahora bien, quizá deberíamos recalcar a los alumnos qué significa cada elemento de la fórmula (puesto que es lo que se pregunta en el ítem 5) y fortalecer así el aprendizaje.

¿Qué propiedades matemáticas debe tener un distractor? Obviamente, tener discriminación negativa, es decir, ser más elegido entre los peores que entre los mejores. Además, sería óptimo que todos los distractores tuvieran una discriminación parecida, ya que indicaría que sus capacidades de atracción son semejantes. Conseguir esto es especialmente difícil, y esta dificultad crece exponencialmente con las alternativas de respuesta. En resumen: es mucho más difícil redactar tres distractores que sean efectivos que dos. Por ello, la mayoría de los estudios realizados recomiendan usar como mucho tres alternativas de respuesta.

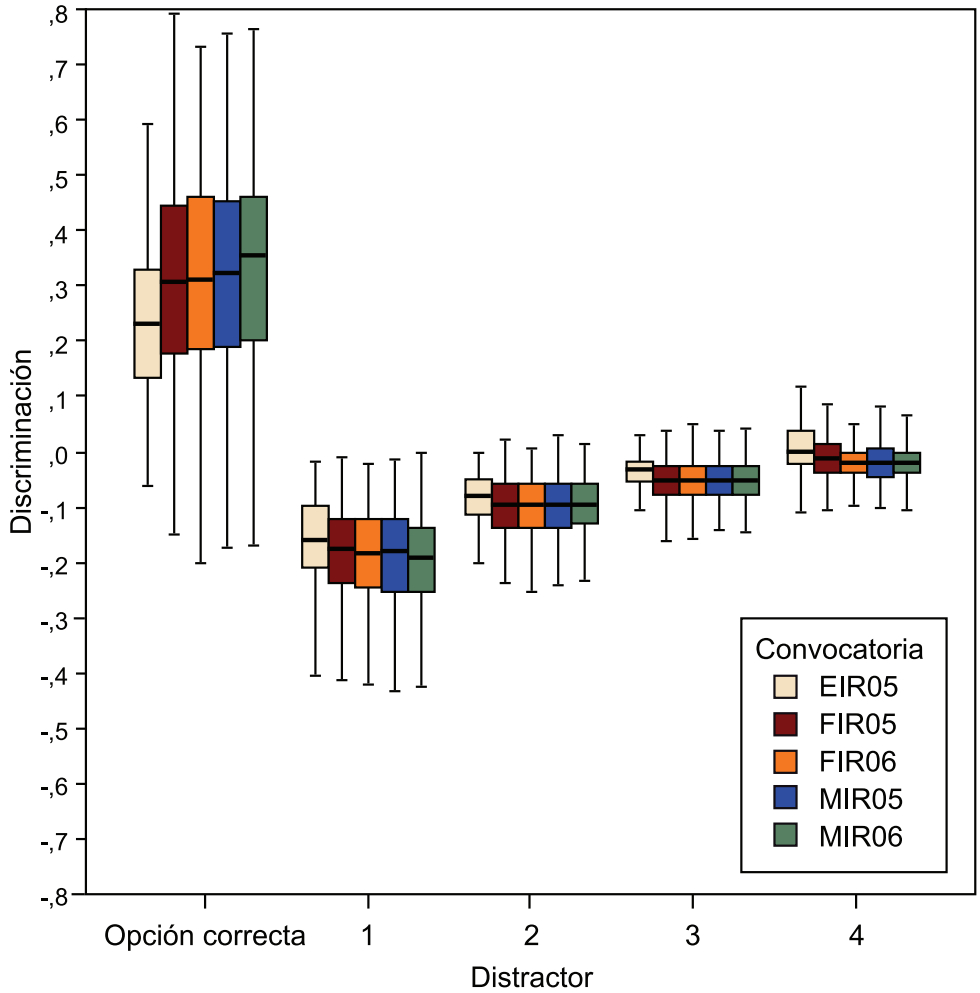
No debemos creer que las propiedades que hasta ahora hemos presentado son independientes entre sí: nada más alejado de la realidad. Si un ítem tiene una opción de respuesta inverosímil (por ejemplo, Maradona como autor de *El Quijote*), el ítem será más fácil y necesariamente discriminará peor.

Como ya hemos hecho cuando hemos estudiado la dificultad, ahora nos planteamos si existe alguna manera de estudiar el conjunto de las discriminaciones de los ítems de una prueba.

De nuevo, la propuesta es nuestra (Bonillo, 2012) y consiste en mostrar en un gráfico, denominado diagrama de cajas, la discriminación de la opción correcta y de cada uno de los distractores, ordenados de mayor a menor.

La figura siguiente muestra esto aplicado a, de nuevo, los exámenes de acceso a la formación sanitaria especializada.

Figura 2. Discriminación de los distractores de las pruebas FSE



Debemos tener en cuenta que en esta figura aparecen las cinco convocatorias analizadas (esto es, [250 ítems × 2 programas × 2 años + 100 ítems de EIR] × 5 alternativas = 5.500 valores). Así, y para cada ítem, el distractor 1 es el más discriminativo, y el 4 el que menos. Como suele ocurrir en estos gráficos, las cajas muestran la media –en trazo grueso– y los cuartos –en los límites de las cajas. Las patillas (*whiskers*) muestran los valores mínimos y máximos no alejados ni extremos. Los alejados se muestran con puntos y los extremos, con asteriscos.

Se observa que las discriminaciones de las alternativas correctas son semejantes entre especialidades y convocatorias. Destacan de las demás las discriminaciones relativas a la prueba de EIR, que son más bajas y menos dispersas. En el análisis de los distractores se observa que existe un escalado entre estos, pero que se reduce cuantas más alternativas se contemplan; es decir, la diferencia entre la tercera y la cuarta alternativa es mucho menor que entre la primera y la segunda. También se observa que las alternativas tres y cuatro –recordemos que son ordenadas por su discriminación y que no deben identificarse con alternativas de respuesta D y E, por ejemplo– tienen discriminaciones muy bajas o casi nulas. Si consideramos que el límite superior de las cajas de la última alternativa es superior a 0, podemos decir que más del 25% de los ítems tienen una alternativa de respuesta con discriminación positiva –es decir, más elegida por el grupo con rendimiento alto. Además, la última alternativa presenta muchos valores extremos y alejados, esto es, ítems en los que, por su alta discriminación positiva, sería discutible si la opción dada como correcta verdaderamente lo es, o es la única que lo es. Conclusión que hay que extraer: con tres opciones sería más que suficiente[5].

3.4. Valoración del sesgo

Un aspecto crucial cuando se crea –y cuando se valora– tanto un ítem como un instrumento de medida es que estos sean no sesgados. ¿De qué hablamos cuando nos referimos al sesgo en un instrumento de medida? Una báscula estará sesgada si siempre infravalora el peso de un objeto frente a otro cuando sabemos que ambos pesan exactamente lo mismo. En el contexto de las pruebas de ejecución máxima, entendemos que un ítem –o un test– está sesgado cuando grupos, por ejemplo hombres y mujeres o ricos y pobres, que tienen el mismo conocimiento sobre la materia medida, no obtienen valores iguales, sino que uno de los grupos es sistemáticamente "perjudicado".

5. De nuevo, si queréis conocer con detalle esta propuesta, podéis consultar el artículo original (Bonillo, 2012), ya que en este capítulo no podemos extendernos mucho más de lo que ya hemos hecho.

Como podéis imaginar, los instrumentos sesgados pueden tener graves implicaciones sociales. Si un examen, como la selectividad, favoreciera sistemáticamente a un alumno con nivel socioeconómico alto frente a uno que no lo tiene, y siempre y cuando sepamos que ambos saben exactamente lo mismo sobre el tema, la selectividad no sería socialmente justa.

¿Cómo valorar el sesgo? Actualmente, se utiliza el concepto de *funcionamiento diferencial de los ítems* y el principal índice que se deriva es el DIF[6]. Decimos que un ítem "tiene o presenta" DIF[7] cuando se dan diferencias estadísticamente significativas en la puntuación de un ítem en dos grupos diferentes que deberían tener, de buena lógica, el mismo nivel. Para evaluar el DIF pueden utilizarse diferentes procedimientos matemáticos[8], pero en este texto solo hablaremos del método de Mantel-Haenszel (1959). Este es, en el marco de TCT, el más utilizado por su sencillez de cálculo y sus buenos resultados[9].

El modo de cálculo y la idea que subyace al DIF son muy sencillos. Supongamos que deseamos saber si nuestra prueba no está afectada por el sexo de los que responden. Es obvio que no debería estarlo, y que si lo está deberíamos corregirlo, ya que podría calificarse de sexista.

Se trata entonces de dividir a los sujetos en grupos en función de sus puntuaciones totales (por ejemplo, en cinco grupos). Luego, habrá que elaborar una tabla por grupo en la que observaremos si la variable sexo se asocia a acertar más. Finalmente, este resultado se agregará en el estadístico de Mantel-Haenszel y se comparará con el χ^2 de referencia. Si el resultado es significativo, es que existen tramos de puntuación total en los que un sexo parece tener ventaja sobre otro, ya que acierta más. Entonces podemos decir que la prueba no es justa. Cuando creemos una prueba deberíamos comprobar que es independiente de variables como el género, la raza y cualquier otra que pueda llevar, implícitamente, a la discriminación de las personas.

6. La sigla DIF proviene de las iniciales del término en inglés.

7. En terminología psicométrica coloquial, decimos que "existe DIF".

8. En Muñiz (2003, pp. 239-253) se puede ver una excelente revisión.

9. Recomendamos encarecidamente la consulta del ejemplo que presenta Muñiz (pp. 247-249) y que aquí describiremos, pero no desarrollaremos matemáticamente.

4. Teoría de respuesta al ítem

La *teoría de respuesta al ítem* (en inglés, *item response theory*, IRT) parte de una perspectiva totalmente distinta. Critica a la TCT al afirmar que estudia las propiedades de un test particular en una muestra –también particular– de personas. Desde la TCT, es cierto que dos tests distintos que miden el mismo constructo tendrán propiedades diferentes. Y también es cierto que esas propiedades dependerán de si las personas utilizadas para calibrar el test de rendimiento tienen o no un alto rendimiento. La TRI permite superar estos problemas, pero a costa de complicar enormemente los cálculos y que resulte más difícil de utilizar al ser más "cajanegrista".

La TRI se basa en el cálculo, para cada ítem, de una serie de parámetros, que asumen un modelo matemático muy concreto. El objetivo fundamental es la medición del rasgo latente[10], a partir de tres parámetros:

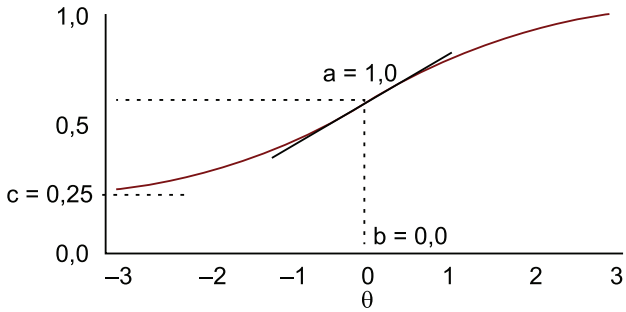
- la discriminación del ítem,
- su dificultad
- y el acierto al azar.

Estos tres parámetros pueden observarse en la figura siguiente, que resulta clave para entender qué es la TRI (en inglés, *item response function*, IRF): se conoce como la curva característica del ítem (CCI).

La CCI muestra, en el eje de las ordenadas (el eje *Y*), la probabilidad de acertar el ítem a partir de la magnitud del rasgo latente (eje abscisas o *X*). Esta probabilidad sigue una función sigmoidal (en forma de *S*, también llamada logística) y el rasgo latente se indica como θ (*theta*). La función logística tiene como propiedad que no puede ser menor de 0 ni mayor de 1, y θ suele estar comprendida entre –3 y +3.

10. El rasgo latente es el nombre que recibe el constructo que se va a medir, por ejemplo, el conocimiento de una asignatura.

Figura 3. Ejemplo de curva característica de un ítem (CCI)



A partir de la CCI podemos medir los tres parámetros clave de la TRI. El parámetro indicado como a) mide la discriminación del ítem. Una curva muy plana expresaría que no es importante tener un alto conocimiento del rasgo para aumentar la probabilidad de acierto. Es decir, cuanto mayor sea la pendiente, mayor será la discriminación. El del gráfico mostrado es 1, valor positivo y aceptable.

El parámetro indicado como *b* mide la dificultad a partir del punto de corte del eje *X*, que corresponde a una probabilidad de acierto del 50%. Se interpretaría como el nivel de rasgo latente necesario para tener un 50% de probabilidades de acertar el ítem. Cuanto mayor sea la *b*, más difícil será el ítem, ya que hará falta más conocimiento para poder llegar a ese 50% de probabilidad deseado. El valor mostrado en el gráfico es 0, que se interpretaría como que es un ítem de dificultad (exactamente) media. El tercer parámetro, indicado como c), mide el nivel de azar y se conoce también como *índice de pseudoadivinación*. Gráficamente, corresponde al valor de *X* que corta el eje *Y*. Contempla, lógicamente, la probabilidad de acertar cuando el conocimiento del ítem es nulo. El valor del gráfico indica que este es alto: del 25%.

Estos cálculos se realizan con software muy específico. Mostrar con qué herramientas hacerlo y cómo excede los objetivos de este capítulo.[11]

11. La siguiente página contiene un conjunto de programas gratuitos que permiten aplicar la TRI: http://www.psychology.gatech.edu/unfolding/FreeSoftware.html. Ahora bien, debemos recordar que aplicar la TRI no es sencillo y que requiere mayor formación que la que necesita la aplicación de la TCT.

Ejemplo de evaluación de las propiedades de los ítems a una muestra de alumnos ficticia

En la siguiente tabla se muestran las respuestas de los veinte sujetos (por filas) a las diez preguntas de la prueba (columnas). En negrita se marcan las respuestas correctas, cuya pauta aparece en la primera fila.

Pauta	**A**	**C**	**B**	**B**	**C**	**B**	**A**	**C**	**C**	**A**
	il	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**
S1	**A**	**C**	C	**B**	**C**	**B**	**A**	**C**	**C**	B
S2	**A**	**C**	C	**B**	B	**B**	**A**	**C**	**C**	**A**
S3	**A**	**C**	C	**B**	B	**B**	**A**	**C**	**C**	**A**
S4	**A**	**C**	C	A	**C**	**B**	**A**	**C**	**C**	**A**
S5	**A**	**C**	A	**B**	B	**B**	**A**	**C**	**C**	**A**
S6	**A**	A	A	A	A	A	B	A	B	B
S7	**A**	**C**	A	**B**	**C**	A	B	B	B	B
S8	**A**	**C**	A	A	V	A	B	B	B	B
S9	**A**	A	A	**B**	**C**	A	B	B	B	C
S10	**A**	**C**	A	**B**	A	A	B	B	B	B
S11	**A**	A	A	**B**	**C**	**B**	B	**C**	**C**	**A**
S12	**A**	A	A	A	A	**B**	B	**C**	**C**	**A**
S13	**A**	A	A	A	**C**	**B**	B	B	B	C
S14	**A**	A	A	A	**C**	**B**	B	B	B	B
S15	**A**	A	A	A	**C**	**B**	B	B	B	C
S16	**A**	A	A	A	B	C	B	B	C	B

Pauta	A	C	B	B	C	B	A	C	C	A
	il	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**
S17	**A**	A	A	A	B	A	B	B	B	B
S18	**A**	A	A	A	A	C	B	A	A	C
S19	**A**	A	A	A	A	A	B	A	A	C
S20	**A**	A	A	A	A	A	B	A	A	C

La tabla reproduce la anterior, pero codificando la respuesta como 1, si ha acertado, y 0, si no lo ha hecho. La columna titulada P. total contiene la puntuación total de cada persona, y corresponde a la suma de los unos de la tabla. En la parte inferior vemos dos filas, denominadas ID e IDc. Como sabemos, la primera corresponde al índice de dificultad y la segunda también a este pero en su versión corregida. Observad la gran diferencia entre ambos y recordad que el segundo corrige el acierto por azar que el primero obvia.

	Acierto										**P. total**
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**	
S1	1	1	0	1	1	1	1	1	1	0	8
S2	1	1	0	1	0	1	1	1	1	1	8
S3	1	1	0	1	0	1	1	1	1	1	8
S4	1	1	0	0	1	1	1	1	1	1	8
S5	1	1	0	1	0	1	1	1	1	1	8
S6	1	0	0	0	0	0	0	0	0	0	1
S7	1	1	0	1	1	0	0	0	0	0	4
S8	1	1	0	0	0	0	0	0	0	0	2

	Acierto										P. total
	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	
S9	1	0	0	1	1	0	0	0	0	0	3
S10	1	1	0	1	0	0	0	0	0	0	3
S11	1	0	0	1	1	1	0	1	1	1	7
S12	1	0	0	0	0	1	0	1	1	1	5
S13	1	0	0	0	1	1	0	0	0	0	3
S14	1	0	0	0	1	1	0	0	0	0	3
S15	1	0	0	0	1	1	0	0	0	0	3
S16	1	0	0	0	0	0	0	0	1	0	2
S17	1	0	0	0	0	0	0	0	0	0	1
S18	1	0	0	0	0	0	0	0	0	0	1
S19	1	0	0	0	0	0	0	0	0	0	1
S20	1	0	0	0	0	0	0	0	0	0	1
ID	100,0%	40,0%	0,0%	40,0%	40,0%	50,0%	25,0%	35,0%	40,0%	30,0%	
IDc	100,0%	10,0%	-50,0%	10,0%	10,0%	25,0%	-12,5%	2,5%	10,0%	-5,0%	

Los puntos de corte de los cuartiles de la puntación total son, respectivamente, 1,75 y 8. Esto es, qué puntuación mínima hay que tener para estar por encima del primer y tercer cuartiles. Los valores comprendidos entre ambos corresponden al grupo denominado intermedio, que contiene el 50% de las personas con valores alrededor de la media. A partir de estos valores se calcula la columna que podemos ver en la derecha de la tabla, y que clasifica a las personas en tres grupos.

	P. total	Grupo Rend.
S1	8	Superior
S2	8	Superior
S3	8	Superior
S4	8	Superior
S5	8	Superior
S6	1	Inferior
S7	4	Intermedio
S8	2	Intermedio
S9	3	Intermedio
S10	3	Intermedio
S11	7	Intermedio
S12	5	Intermedio
S13	3	Intermedio
S14	3	Intermedio
S15	3	Intermedio
S16	2	Intermedio
S17	1	Inferior
S18	1	Inferior
S19	1	Inferior
S20	1	Inferior

Las celdas de la tabla siguiente muestran los cálculos relativos a la discriminación de cada uno de los ítems. En las celdas tituladas % acierto grupo superior se muestra el porcentaje de personas del grupo de rendimiento superior que han elegido, respectivamente, las alternativas A, B y C. Idéntico cálculo se realiza en las celdas tituladas % acierto del grupo inferior. ¿Qué hacemos posteriormente con estos valores

Gracias a este cálculo podemos obtener de manera sencilla la discriminación de cada una de las opciones de respuesta. En las celdas tituladas Índice D alternativas vemos la discriminación de la alternativa A, y lo mismo para las B y C con las filas inferiores. En negrita se muestra la discriminación de la alternativa correcta. Observad que la negrita coincide con la letra que, en la primera tabla de este apartado, aparecía en la pauta de respuestas correctas.

	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
% acierto grupo superior										
A	100,0%	0,0%	20,0%	20,0%	0,0%	0,0%	100,0%	0,0%	0,0%	80,0%
B	0,0%	0,0%	0,0%	80,0%	60,0%	100,0%	0,0%	0,0%	0,0%	20,0%
C	0,0%	100,0%	80,0%	0,0%	40,0%	0,0%	0,0%	100,0%	100,0%	0,0%
% acierto grupo inferior										
A	100,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
B	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
C	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Índice D alternativas										
A	0,0%	0,0%	20,0%	20,0%	0,0%	0,0%	100,0%	0,0%	0,0%	80,0%
B	0,0%	0,0%	0,0%	80,0%	60,0%	100,0%	0,0%	0,0%	0,0%	20,0%
C	0,0%	100,0%	80,0%	0,0%	40,0%	0,0%	0,0%	100,0%	100,0%	0,0%

5. Conclusiones

A lo largo de este texto hemos expuesto una multitud de aspectos sobre los instrumentos de medida. Sobre una prueba hemos visto qué estudiar, cómo y cuándo hacerlo, y hemos sido –especial y conscientemente– insistentes en el porqué hacerlo. La idea principal que nos gustaría haber podido transmitir es que lo que no podemos hacer es no hacer nada. Es decir, la peor de las situaciones posibles que podemos imaginar –en este contexto, claro– es aplicar una prueba sin estudiar ninguna de sus propiedades. No hacerlo convertirá el error en norma y no en excepción.

Esta situación descrita es, lamentablemente, la realidad. En el ámbito universitario, que es el que nos resulta más cercano, son contados los profesores que estudian sus exámenes tras administrarlos. Incluso los que lo hacen rara vez publican los resultados, lo que permitiría a los alumnos contrastar la justicia de la prueba con la que se les examinó. ¿Por qué ocurre esto? Creemos que es más achacable a la falta de formación que a la falta de transparencia. Al profesor no se le forma en cómo debe valorar sus pruebas, como tampoco se le forma en cómo debe realizar una clase. En nuestro sistema educativo, y estamos refiriéndonos a todas las etapas, la cultura de trabajar con evidencias no está implantada.

No podemos no centrarnos en el ámbito educativo, que es al que pertenecemos, pero ¿no creéis que deberían publicarse los datos que permitan valorar la justicia de unas oposiciones, como son las pruebas de acceso de la formación sanitaria especializada? De hecho, en el artículo ya citado (Bonillo, 2012) esta fue nuestra principal propuesta. Así, los opositores podrían impugnar preguntas, proponer otras correcciones y, en definitiva, todos podríamos estar más tranquilos al saber que las pruebas son justas y premian de verdad a los mejores.

En este capítulo se han presentado muchos conceptos novedosos y se ha realizado muy brevemente. ¿Consideramos que el lector está preparado para aplicarlos mañana? En absoluto. Nos gustaría pensar que, para aquel que necesite un día crear y/o valorar una prueba, este texto es el primero de otros. Esos otros están citados o pueden buscarse alternativas.

Para aquel lector que no necesite lo que aquí se expone, que será la gran mayoría, confiamos en que los conceptos expuestos se entiendan. Cuando esto

ocurre, y en el futuro se necesita aplicarlos, recuperarlos de la memoria y de la biblioteca es sencillo.

En cualquier caso, y para todos, esperamos haber insuflado espíritu crítico y deseo de trabajar con y por las evidencias. Al final, estos dos elementos son los que separan la psicología de otras cosas que nos deben ser ajenas.

Tablas de distribución

Julio Meneses

Tabla 1. Distribución normal para $x = x_1 + x_2$, $P(Z < z)$

$x_1 \backslash x_2$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.5	0.0002	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.4	0.0003	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005
-3.3	0.0005	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007
-3.2	0.0007	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
-3.1	0.0010	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013
-3.0	0.0013	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
-2.9	0.0019	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
-2.8	0.0026	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
-2.7	0.0035	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
-2.6	0.0047	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
-2.5	0.0062	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080
-2.4	0.0082	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
-2.3	0.0107	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
-2.2	0.0139	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
-2.1	0.0179	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
-2.0	0.0228	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
-1.9	0.0287	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
-1.8	0.0359	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
-1.7	0.0446	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537

$x_1 \backslash x_2$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.6	0.0548	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655
-1.5	0.0668	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793
-1.4	0.0808	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951
-1.3	0.0968	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131
-1.2	0.1151	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335
-1.1	0.1357	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562
-1.0	0.1587	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814
-0.9	0.1841	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090
-0.8	0.2119	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389
-0.7	0.2420	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709
-0.6	0.2743	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050
-0.5	0.3085	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409
-0.4	0.3446	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783
-0.3	0.3821	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168
-0.2	0.4207	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562
-0.1	0.4602	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

$x_1 \backslash x_2$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

Tabla 2. Distribución χ^2 de Pearson, $P(X \leq \chi^2_{p,n})$

n\α	0.005	0.010	0.025	0.050	0.100	0.500	0.900	0.950	0.975	0.990	0.995
1	0.000	0.000	0.001	0.004	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	4.351	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	11.340	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	13.339	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	14.339	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	15.338	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	16.338	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	17.338	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	18.338	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	19.337	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	20.337	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	21.337	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	22.337	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	23.337	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	24.337	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	25.336	35.563	38.885	41.923	45.642	48.290

n\α	0.005	0.010	0.025	0.050	0.100	0.500	0.900	0.950	0.975	0.990	0.995
27	11.808	12.879	14.573	16.151	18.114	26.336	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	27.336	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	28.336	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	29.336	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	39.335	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	49.335	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	59.335	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	69.334	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	79.334	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	89.334	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	99.334	118.498	124.342	129.561	135.807	140.169

Tabla 3. Distribución *t* de Student, $P(T \leq t_{p,n})$

n\α	0.550	0.600	0.650	0.700	0.750	0.800	0.850	0.900	0.950	0.975	0.990	0.995	0.999
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.309
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421

n\α	0.550	0.600	0.650	0.700	0.750	0.800	0.850	0.900	0.950	0.975	0.990	0.995	0.999
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307
50	0.126	0.255	0.388	0.528	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261
60	0.126	0.254	0.387	0.527	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232
80	0.126	0.254	0.387	0.526	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195
100	0.126	0.254	0.386	0.526	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.160
200	0.126	0.254	0.386	0.525	0.676	0.843	1.039	1.286	1.653	1.972	2.345	2.601	3.131
500	0.126	0.253	0.386	0.525	0.675	0.842	1.038	1.283	1.648	1.965	2.334	2.586	3.107
1000	0.126	0.253	0.385	0.525	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090

Tabla 4. Distribución F de Snedecor para $\alpha = 0.900$, $P(F_{n_1,n_2} \leq f_{\alpha,n_1,n_2})$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	50	60	100	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.69	62.79	63.0	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.481	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.13	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.77	2.76	2.75	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.52	2.51	2.50	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.22	2.21	2.19	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.12	2.11	2.09	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.01	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.97	1.96	1.94	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.87	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.83	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.79	1.78	1.76	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.76	1.75	1.73	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.74	1.72	1.70	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.71	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.69	1.68	1.65	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.67	1.66	1.63	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.61	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.64	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.58	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.61	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.59	1.58	1.55	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.58	1.57	1.54	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.57	1.56	1.53	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.56	1.55	1.52	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.55	1.54	1.51	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.43	1.38
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.68	1.63	1.57	1.54	1.50	1.46	1.44	1.42	1.39	1.33
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.41	1.40	1.36	1.29
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.61	1.56	1.49	1.46	1.42	1.38	1.35	1.34	1.29	1.21
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.58	1.52	1.46	1.42	1.38	1.34	1.31	1.29	1.24	1.14
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.26	1.24	1.18	1.00

Tabla 5. Distribución F de Snedecor para $\alpha = 0.950$, $P\left(F_{n_1,n_2} \leq f_{\alpha,n_1,n_2}\right)$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	50	60	100	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	251.77	252.20	253.04	254.31
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	19.48	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	8.58	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	5.70	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	4.44	4.43	4.41	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	3.75	3.74	3.71	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	3.32	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	3.02	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	2.80	2.79	2.76	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	2.64	2.62	2.59	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	2.51	2.49	2.46	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	2.40	2.38	2.35	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	2.31	2.30	2.26	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	2.24	2.22	2.19	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	2.18	2.16	2.12	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	2.12	2.11	2.07	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	2.08	2.06	2.02	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	2.04	2.02	1.98	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	2.00	1.98	1.94	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	1.97	1.95	1.91	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	1.94	1.92	1.88	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	1.91	1.89	1.85	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	1.88	1.86	1.82	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.86	1.84	1.80	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	1.84	1.82	1.78	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	1.82	1.80	1.76	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	1.81	1.79	1.74	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	1.79	1.77	1.73	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	1.77	1.75	1.71	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	1.76	1.74	1.70	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	1.66	1.64	1.59	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.74	1.69	1.63	1.60	1.58	1.52	1.44
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	1.56	1.53	1.48	1.39
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.77	1.68	1.63	1.57	1.52	1.48	1.45	1.39	1.28
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.72	1.62	1.57	1.52	1.46	1.41	1.39	1.32	1.19
∞	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	1.35	1.32	1.24	1.00

Tabla 6. Distribución *F* de Snedecor para $\alpha = 0.975$, $P\left(F_{n_1,n_2} \leq f_{\alpha,n_1,n_2}\right)$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	50	60	100	∞
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	976.71	984.87	993.10	997.25	1001.41	1005.60	1008.12	1009.80	1013.17	1018.26
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	39.48	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	14.01	13.99	13.96	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	8.38	8.36	8.32	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	6.14	6.12	6.08	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	4.98	4.96	4.92	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	4.41	4.36	4.31	4.28	4.25	4.21	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	3.81	3.78	3.74	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	3.47	3.45	3.40	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	3.22	3.20	3.15	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	3.03	3.00	2.96	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	2.87	2.85	2.80	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	2.74	2.72	2.67	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	2.64	2.61	2.56	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	2.55	2.52	2.47	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	2.47	2.45	2.40	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	2.41	2.38	2.33	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	2.35	2.32	2.27	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	2.30	2.27	2.22	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	2.25	2.22	2.17	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.21	2.18	2.13	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	2.17	2.14	2.09	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	2.14	2.11	2.06	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	2.11	2.08	2.02	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	2.08	2.05	2.00	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	2.05	2.03	1.97	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	2.03	2.00	1.94	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	2.01	1.98	1.92	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	1.99	1.96	1.90	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	1.97	1.94	1.88	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	1.83	1.80	1.74	1.64
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.22	2.11	1.99	1.93	1.87	1.80	1.75	1.72	1.66	1.55
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	1.70	1.67	1.60	1.48
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.08	1.97	1.85	1.78	1.71	1.64	1.59	1.56	1.48	1.35
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	2.01	1.90	1.78	1.71	1.64	1.56	1.51	1.47	1.39	1.23
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.43	1.39	1.30	1.00

Tabla 7. Distribución *F* de Snedecor para $\alpha = 0.990$, $P\left(F_{n_1,n_2} \leq f_{\alpha,n_1,n_2}\right)$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	50	60	100	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6303	6313	6334	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	99.48	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	26.35	26.32	26.24	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	13.69	13.65	13.58	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	9.24	9.20	9.13	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	7.09	7.06	6.99	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	5.86	5.82	5.75	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	5.07	5.03	4.96	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	4.81	4.73	4.65	4.57	4.52	4.48	4.41	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	4.12	4.08	4.01	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	3.81	3.78	3.71	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	3.57	3.54	3.47	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	3.38	3.34	3.27	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	3.22	3.18	3.11	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	3.08	3.05	2.98	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	2.97	2.93	2.86	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	2.87	2.83	2.76	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	2.78	2.75	2.68	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	2.71	2.67	2.60	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	2.64	2.61	2.54	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	2.58	2.55	2.48	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	2.53	2.50	2.42	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	2.48	2.45	2.37	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	2.44	2.40	2.33	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	2.40	2.36	2.29	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	2.36	2.33	2.25	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	2.33	2.29	2.22	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	2.30	2.26	2.19	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	2.73	2.57	2.49	2.41	2.33	2.27	2.23	2.16	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	2.25	2.21	2.13	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	2.06	2.02	1.94	1.80
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.27	2.18	2.10	2.01	1.95	1.91	1.82	1.68
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	1.88	1.84	1.75	1.60
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.22	2.07	1.98	1.89	1.80	1.74	1.69	1.60	1.43
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.13	1.97	1.89	1.79	1.69	1.63	1.58	1.48	1.28
∞	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	1.52	1.47	1.36	1.00

Tabla 8. Distribución *F* de Snedecor para $\alpha = 0.995$, $P\left(F_{n_1,n_2} \leq f_{\alpha,n_1,n_2}\right)$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	50	60	100	∞
1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	24940	25044	25148	25211	25253	25337	25464
2	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.37	199.39	199.40	199.42	199.43	199.45	199.46	199.47	199.47	199.48	199.48	199.49	199.50
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	42.21	42.15	42.02	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	19.67	19.61	19.50	19.32
5	22.78	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	12.45	12.40	12.30	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	9.17	9.12	9.03	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	7.64	7.53	7.42	7.35	7.31	7.22	7.08
8	14.69	11.04	9.60	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	6.22	6.18	6.09	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	5.45	5.41	5.32	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	4.90	4.86	4.77	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	4.49	4.45	4.36	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	4.17	4.12	4.04	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	3.91	3.87	3.78	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	3.70	3.66	3.57	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	3.52	3.48	3.39	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	3.37	3.33	3.25	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	3.25	3.21	3.12	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	3.14	3.10	3.01	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	3.04	3.00	2.91	2.78
20	9.94	6.99	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	2.96	2.92	2.83	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	2.88	2.84	2.75	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	2.82	2.77	2.69	2.55
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.76	2.71	2.62	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	2.70	2.66	2.57	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	2.65	2.61	2.52	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	2.61	2.56	2.47	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	2.57	2.52	2.43	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	2.53	2.48	2.39	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	2.49	2.45	2.36	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	2.46	2.42	2.32	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	2.23	2.18	2.09	1.93
50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.82	2.65	2.47	2.37	2.27	2.16	2.10	2.05	1.95	1.79
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	2.01	1.96	1.86	1.69
100	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	2.58	2.41	2.23	2.13	2.02	1.91	1.84	1.79	1.68	1.49
200	8.06	5.44	4.41	3.84	3.47	3.21	3.01	2.86	2.73	2.63	2.47	2.30	2.11	2.01	1.91	1.79	1.71	1.66	1.54	1.31
∞	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	1.59	1.53	1.40	1.00

Bibliografía

Capítulo I. Aproximación histórica y conceptos básicos de la psicometría

American Educational Research Association, American Psychological Association, y National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association (2010). *Ethical principles of psychologists and code of conduct*. Washington, DC: American Psychological Association. Disponible en línea en: http://www.apa.org/ethics/code/principles.pdf.

Anastasi, A. (1988). *Psychological testing* (6.ª ed.). New York: Macmillan.

Bermejo, V. (2009). Motivaciones para la revisión y cambios en el código deontológico de la profesión de la psicología. *Papeles del Psicólogo*, 30 (3), 195-206. Disponible en línea en: http://www.papelesdelpsicologo.es/pdf/1748.pdf

Bock, R. D. (1997). A brief history of Item response theory. *Educational Measurement: Issues and Practice*, 16 (4), 21-33.

Bock, R. D. y Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Bondy, M. (1974). Psychiatric antecedents of psychological testing (before Binet). *Journal of the History of the Behavioral Sciences*, 10 (2), 180-194.

Boring, E. G. (1978). *Historia de la psicología experimental*. México: Trillas.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in moderno psychometrics*. Cambridge: Cambridge University Press.

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist*, 44 (3), 576-578.

Brennan, R. L. (1997). A Perspective on the history of Generalizability Theory. *Educational Measurement: Issues and Practice*, 16 (4), 14-20.

Bridgman, P. W. (1927). *The logic of modem physics*. New York: Macmillan.

Buchanan, R. D. y Finch, S. J. (2005). *History of psychometrics*. En B. S. Everitt y D. Howell (Ed.), *Encyclopedia of statistics in behavioral science* (pp. 875-878). Chichester: Wiley.

Campbell, L., Vasquez, M., Behne, S., y Kinscherff, R. (2010). *APA Ethics code commentary and caso illustrations*. Washington, DC: American Psychological Association.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge: Cambridge University Press.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London: Longmans, Green, and co.

Carson, J. (1993). Army alpha, army brass, and the search for army intelligence. *Isis*, 84 (2), 278-309.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological bulletin*, 40 (3), 153-193.

Chadha, N. K. (2009). *Applied psychometry*. New Delhi: Sage.

Consejo General de Colegios Oficiales de Psicólogos (2010). *Código deontológico*. Madrid: Consejo General de Colegios Oficiales de Psicólogos. Disponible en línea en: http://www.cop.es/pdf/codigo-deontologico-consejo-adaptacion-ley-omnibus.pdf.

Crocker, L. y Algina, J. (2006). *Introduction to Classical and Modern test theory*. Mason: CENGAGE Learning.

Cronbach, L. J., Rajaratnam, N., y Gleser, G. C. (1963). Theory of Generalizability: A liberation of reliability theory. *British Journal of Mathematical and Statistical Psychology*, 16 (2), 137-163.

Dingle, H. (1950). A theory of measurement. *The British Journal for the Philosophy of Science*, 1 (1), 5-26.

Downing, S. M. (2006). *Twelve steps for effective test development*. En S. M. Downing y T. M. Haladyna (Ed.), *Handbook of test development* (pp. 3-25). Mahwah: Lawrence Erlbaum.

DuBois, P. H. (1970). *The history of psychological testing*. Boston: Allyn & Bacon.

European Federation of Psychologists' Associations (2005). *Meta-Code of Ethics*. Brussels: European Federation of Psychologists' Associations. Disponible en línea en: http://www.efpa.eu/ethics/ethical-codes.

European Federation of Psychologists' Associations (2012). *EFPA Review modelo for the description and evaluation of psychological and educational tests*. Versión 4.2.4. Brussels: European Federation of Psychologists' Associations. Disponible en línea en: http://www.efpa.eu/professional-development.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., Campbell, N. R., Craik, K. J. W., Drever, J., Guild, J., Houstoun, R. A., Irwin, J. O., Kaye, G. W. C., Philpott, S. J. F., Richardson, L. F., Shaxby, J. H., Smith, T., Thouless, R. H., y Tucker, W. S. (1940). *Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events*. *Advancement of Science*, 1, 331-349.

Fernández-Ballesteros, R. (1997). Evaluación psicológica y tests. En A. Cordero (Ed.), *La evaluación psicológica en el año 2000* (pp. 11-26). Madrid: TEA Ediciones.

Fraser, C. O. (1980). Measurement in psychology. *British Journal of Psychology*, 71 (1), 23-34.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *History of Psychology*, 87 (3), 564-567.

García Cueto, E. (1993). *Introducción a la psicometría*. Madrid: Siglo Veintiuno de España Editoras.

Gibby, R. E. y Zickar, M. J. (2008). A history of the early days of personality testing in American industry: An obsession with adjustment. *History of Psychology*, 11 (3), 164-184.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, 18 (8), 519-521.

Gleser, G. C., Cronbach, L. J., y Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30 (4), 395-418.

Goodenough, F. L. (1949). *Mental testing: Its history, principles, and applications*. New York: Rinehart.

Goslin, D. A. (1963). *The search for ability. Standardized testing in social perspective*. New York: Russell Sage Foundation.

de Gruijter, D. N. M. y van der Kamp, L. J. T. (2008). *Statistical test theory for the Behavioral Sciences*. Boca Raton: Chapman & Hall.

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Willey.

Gulliksen, H. (1968). Louis Leon Thurstone, experimental and mathematical psychologist. *History of Psychology*, 23 (11), 786-802.

Hambleton, R. K. (1994). The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice*, 13 (4), 21-26.

Hambleton, R. K., Swaminathan, H., y Rogers, H. J. (1991). *Fundamentals of Item response theory*. Newbury Park: Sage Publications.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159 (3), 445-492.

Hilgard, E. R. (1987). *Psychology in America: A historical survey*. Orlando: Harcourt.

Holden, R. R. (2000). *Psychometrics*. En A. E. Kazdin (Ed.), *Encyclopedia of psychology*, VI (pp. 417-419). New York: Oxford University Press.

International Test Commission (2000). *International guidelines for test use*. Disponible en línea en: http://www.intestcom.org/upload/sitefiles/41.pdf.

Kline, P. (1998). *The new psychometrics*. Science, psychology, and measurement. London: Routledge.

Jáñez, L. (1989). *Fundamentos de psicología matemática*. Madrid: Pirámide.

Jensen, A. R. (1980). *Bias in mental testing* . New York: Free Press.

Jones, L. V. (1971). *The nature of measurement*. En R. L. Thorndike (Ed.), *Educational measurement* (2.ª ed.) (pp. 335-355). Washington, DC: American Council donde Education.

Jones, L. V. y Thissen, D. (2007). *A history and overview of Psychometrics*. En C. R. Rao y S. Sinharay (Ed.), *Handbook of statistics 26. Psychometrics* (pp. 1-27). Amsterdam: Elsevier.

Kaplan, R. M. y Saccuzzo, D. P. (2001). *Psychological testing* (6.ª ed.). Belmont: Wadsworth Publishing Company.

Lord, F. M. (1953). On the statistical treatment of football numbers. *The American Psychologist*, 8 (12), 750-751.

Lord, F. M. y Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison Wesley.

Martin, O. (1997). La mesure en psychologie de Binet à Thurstone, 1900-1930. *Revue de synthèse*, 118 (4), 457-493.

Martínez Arias, M. R., Hernández Lloreda, M. J., y Hernández Lloreda, M. V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Maydeu-Olivares, A. y McArdle, J. J. (2005). *Contemporary psychometrics*. Mahwah: Lawrence Erlbaum.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *History of Psychology*, 100 (3), 398-407.

Michell, J. (1999). *Measurement in psychology*. Cambridge: Cambridge University Press.

Muñiz, J. (1996). *Psicometría*. Madrid: Editorial Universitas.

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.

Muñiz, J. (2010). Las teorías de los tests: Teoría clásica y Teoría de respuesta a los ítems. P*apeles del Psicólogo*, 31 (1), 57-66. Disponible en línea en: http://www.papelesdelpsicologo.es/pdf/1796.pdf.

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, A., y Peña-Suárez, E. (2011). Evaluación de tests editados en España. *Papeles del Psicólogo*, 32 (2), 113-128. Disponible en línea en: http://www.papelesdelpsicologo.es/pdf/1947.pdf.

Muñiz, J. y Fonseca-Pedrero, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación*, 5, 13-25.

Muñiz, J. y Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de Psicología*, (52), 41-66.

Murphy, K. R. y Davidshofer, C. O. (2005). *Psychological testing* (6.ª ed.). Upper Saddle River: Pearson education.

Nicolas, S. y Ferrand, L. (2002). Alfred Binet and higher education. *History of Psychology*, 5 (3), 264-283.

Padilla, M., Merino, J. M., Rodríguez-Miñón, P., y Moreno, E. (1996). *Psicología matemática I*. Madrid: UNED.

Popham, W. J. y Husek, T. R. (1969). Implications of criterion-referenced measurement, *Journal of Educational Measurement*, 6 (1), 1-9.

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, 77, 65-75. Disponible en línea en: http://www.papelesdelpsicologo.es/vernumero.asp?id=1102.

Rogers, T. B. (1995). *The psychological testing enterprise*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Psychometrika*, 16 (2), 170-233.

Rust, J. y Golombok, S. (2009). *Modern psychometrics. The science of psychological assessment* (3.ª Ed.). London: Routledge.

Samejima, F. (2000). Psychometric Society. En A. E. Kazdin (Ed.), *Encyclopedia of psychology*, VI (pp. 419-420). New York: Oxford University Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 667-680.

Sokal, M. M. (1982). *James McKeen Cattell and the failure of anthropometric mental testing, 1890-1901*. En W. R. Woodward y M. G. Ash (Ed.), *The problematic science. Psychology in nineteenth-century thought* (pp. 322-345). New York: Praeger Publishers.

Suppes, P. (1951). A set of independiente axioms for extensive quantities. *PortugaliaeMathematica*, 10, 163-172.

Swistak, P. (1990). Paradigms of measurement. *Theory and Decision*, 29 (1), 1-17.

Thurstone, L. L. (1931). *The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems*. Ann Arbor: Edwards Brothers.

Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of the vectores of the mind*. Chicago: Chicago University Press.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16 (4), 8-14.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Urbina, S. (2004). *Essentials of psychological testing*. Hoboken: Wiley.

Valentine, E. R. (1999). The founding of the psychological laboratory, University College London: "Dear Galton…Yours truly, J Sully". *History of Psychology*, 2 (3), 204-218.

Velleman, P. F. y Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, 47 (1), 65-72.

Wolf, T. H. (1969).The emergence of Binet's conception and measurement of intelligence: A case history of the creative process. *Journal of the History of the Behavioral Sciences*, 5 (2), 113-134.

Yela, M. (1984) *Introducción a la teoría de los tests*. Madrid: Universidad Complutense.

Zeidner, J. y Drucker, A. J. (1988). *Behavioral science in the Army: A corporate history of the Army Research Institute*. Alejandría: United States Army Research Institute for the Behavioral and Social Sciences.

Capítulo II. Fiabilidad

Angoff, W. H. (1971). Scales norms and equivalent scores. En R. L. Thorndike (Ed.), *Educational measurement* (2.ª ed.). Washington: American Council on Education.

Angoff, W. H. (1984). *Scales norms and equivalent scores*. Princeton: NJ. Educational Testing Service.

Barbero, M. I., Vila, E., y Suárez, L. C. (2003). *Psicometría*. Madrid: UNED.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45 (2), 4-9.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced test. *Review of Educational Research*, 56 (1), 137-172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, 21 (2), 147-152.

Brennan, R. L. y Wan, L. (2004). A bootstrap procedure for estimating decision consistency for single-administration complex assessments. (CASMA Research Report, núm. 7). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Brennan, R., L. y Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14 (3), 277-289.

Breyer, F. J., y Lewis, C. (1994). Pass-fail reliability for tests with cut scores: A simplified method. *ETS Research Report*, 94-39. Princeton, NJ: Educational Testing Service.

Cizek, G. J. y Bunch, M. B. (2007). *Standard setting*. A guide to establishing and evaluating performance standards on tests. Thoundand Oaks: SAGE Publications, Inc.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37-46.

Cohen, R. J. y Swerdlik, M. E. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7.ª ed.). Boston: McGraw-Hill.

Crocker L. y Algina, J. (1986). *Introduction to classical modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alfa and the internal structures of tests. *Psychometrika*, 16 (3), 297-334.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 30 (3), 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 tests. *Psychometrika*, 34 (3), 363-373.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*, 45 (1), 99-105.

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 (4), 529.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psichometrika*, 10 (4), 255-282.

Hambleton, R. K. y Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced test. *Journal of Educational Measurement*, 10 (3), 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 (1), 41-55.

Hofstee,W. K. (1983). The case for compromise in educational selection and grading. En S. B. Anderson, S. B. y J. S. Helmick (Eds.), *On educational testing*. San Francisco: Jossey-Bass.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13 (4), 253-264.

Jaeger, R. M. (1989). Certification of student competence. En R. L. Linn (Ed.), *Educational measurement* (3.ª ed.). New York: Macmillan.

Kapplan, R. M. y Saccuzo, D. P. (2009). *Psychological testing. Principles, applications and issues* (7.ª ed.). Edition. Belmont, CA: Wadsworth.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. Psychometrika, 28 (3), 221-238.

Kuder, G. F. y Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2 (3), 151-160.

Landis, J. R. y Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159-74.

Lee, W. (2005). Classification consistency under the compound multinomial model. (CASMA Research Report, núm. 13). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Lee, W., Brennan, R. L., y Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement*, 33 (5), 374-390.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45 (1), 255-268.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 9 (1), 13-26.

Livingston, S. A. y Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32 (2), 179-197.

Martínez-Arias, R. (2010). La evaluación del desempeño. *Papeles del psicólogo*, 31 (1), 85-96.

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide

Murphy, K. R. y Davidshofer, C. O. (2005). *Psychological testing. Principles and applications* (6.ª ed.). Upper Saddle River, NJ: Prentice Hall.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14 (1), 3-19.

Nunnally, J. C. (1978). *Psychometric theory* (2.ª ed.). New York: McGraw-Hill.

Plake, B. S. y Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. En G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99-103.

Shrock, S. A. y Coscarelli, W. C. (2007). Criterion-referenced test development: Technical and legal guidelines for corporate training and certification (3.ª ed.). San Francisco, CA: John Wiley and Sons.

Sireci, S. G., Hambleton, R. K., y Pitoniak, M. J. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Review*, 15 (1), 21-25.

Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13 (4), 265-275.

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.

Zieky, M. J. y Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests. Princeton*, NJ: Educational Testing Service.

Capítulo III. Validez

AERA, A. N. (1974). *Standards for educational and psychological tests*. Whashington, DC: American Psychological Association.

AERA, APA, y NCME (1966). *Standards for educational and psychological test and manuals*. Washington, DC: AERA.

AERA, APA, y NCME (1999). *Standards for educational and psychological testing*. Washington DC: AERA.

Anastasi, A. (1954). *Psychological testing*. New York: Macmillan.

Beck, A., Rush, A. J., Shawn, B. F., y Emery, G. (1979).*Cognitive therapy of depression*. New York: Guilford Press.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. Oxford, England: Harpers.

Campbell, D. T. y Fiske, A. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike. *Educational measurement* (pp. 443-507). Washington DC: American Council on Education.

Cronbach, L. J. y Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Czaja, R. y Blair, J. (1996). *Designing Surveys*. Thousand Oaks, CA: Sage Publications.

Ebel, R. (1961). Must all tests be valid? *American Psychologist*, 16, 640-647.

Guilford, J. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, pp. 427-439.

Hamilton, M. (1960). A rating scale for depression. J*ournal Neurol. Neurosurg. Psychiatry*, 23, 56-62.

Kane, M. (2006). Content-Related Validity Evidence in Test Development. En S. M. Downing y T. M. Haladyna. *Handbook of test development* (pp. 131-153). New Jersey: Erlbaum.

Leighton, J. P. (2004). Avoiding misconception, missed use, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6-15.

Messick, S. (1989). Validity. En R. Linn. *Educational Measurement*. 3.ª ed. (pp. 13-104). New York: Macmillan.

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.

Prieto, G. y Delgado, A. R. (2010). Fiabilidad y Validez. *Papeles del Psicólogo*, 31, 67-74.

Ramos-Brieva, J. C. (1986). Validación de la versión castellana de la escala de Hamilton para la depresión. *Actas Luso-Esp. Neurol. Psiquiatr.*, 22, 21-28.

Scott, W. (1917). A fourth method of checking results in vocational selection. *Journal of Applied Psychology*, pp. 61-66.

Shepard, L. (1993). Evaluating test validity. En L. Darling-Hammond. *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16 (2), 5-13.

Shepard, L. A., Camilli, G., Linn, R., y Bohrnstedt, G. (1993). *Setting performance standards for achievement tests*. Stanford, CA: National Academy of Education.

Sireci, S. (1998). The construct of content validity. *Social Indicators Research*, 45, 83-117.

Sireci, S. G. (1998). Gathering and evaluating content validity data. *Educational Assessment*, 5 (4), 299-321.

Spearman, C. (1907). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.

Capítulo IV. Transformación e interpretación de las puntuaciones

Aiken, L. R. (1996). *Tests psicológicos y evaluación* (8.ª ed.) México: Prentice Hall.

Alegret, M., Espinosa, A., Vinyes-Junqué, G., Valero, S., Hernández, I., Tárraga, L. *et al*. (2012). Normative data of a brief neuropsychological battery for Spanish individuals older than 49. *Journal of Clinical and Experimental Neuropsychology*, 34, 209-219.

Aluja, A., Blanch, A., Solé, D., Dolcet J.M., y Gallart, S. (2008). Validez convergente y estructural del NEO-PI-R. Baremos orientativos. *Boletín de Psicología*, 92, 7-25.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washintong, DC: American Psychological Association.

Gomà-i-Freixanet, M. y Valero, S. (2008). Spanish normative data of the Zuckerman-Kuhlman Personality Questionnaire in a general population sample. *Psicothema*, 20, 324-330.

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.

Ramos-Quiroga, J. A ., Draigue, C., Valero, S., Bosch, R., Gómez-Barros, N., Nogueira, M., Palomar, G., Roncero, C., y Casas, M. (2009). Validación al español de la escala de cribado del trastorno por déficit de Atención/hiperactividad en adultos (ASRS v. 1.1): una nueva estrategia de puntuación. *Revista de Neurología*, 48, 449-452.

Rodríguez, C., Jiménez, J. E., Díaz, A., García, E., Martin, R., y Fernández, S. (2012). Datos normativos para el Test de los Cinco Dígitos: desarrollo evolutivo de la flexibilidad en Educación Primaria. *European Journal of Education and Psychology*, 5, 27-38.

Valero, S., Ramos-Quiroga, J. A., Gomà-i-Freixanet, M., Bosch, R., Gómez-Barros, N., Nogueira, M., Palomar, G. *et al*. (2012). Personality profile of adult ADHD: The alternative five factor model. *Personality Research*, 198, 130-134.

Yueh-Hsien, L., Chwen-Yng, S., Wei-Yuan, G., y Yee-Pay, W. (2012). Psychometric validation and normative data of the second Chinese version of the Hooper Visual Organization test in Children. *Research in Developmental Disabilities*, 33, 1919-1927.

Zuckerman, M., Kuhlman, D. M., Joireman, J., Teta, P., y Kraft, M. 1993 A comparison of the three structural models for personality: the big three, the big five, and the alternative five. *Journal of Personality and Social Psychology*, 65, 757-768.

Capítulo V. Análisis de los ítems

Baker, F. (2001). *The Basics of Item Response Theory*. *ERIC Clearinghouse on Assessment and Evaluation*. Maryland: University of Maryland.

Bonillo, A. (2012). Pruebas de acceso a la formación sanitaria especializada para médicos y otros profesionales sanitarios en España: examinando el examen y los examinados. *Gaceta Sanitaria*, 26 (3), pág. 231-235

Downing S. M. (2005). The effects of violating standard item writing principles on test and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10, 133-143

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood: Prentice-Hall.

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 (3), 309-334.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30 (1), 17-24. Disponible en: 10.1037/h0057123.

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.

Martínez Arias, R. (1996). *Psicometría: Teoría de los Tests Psicológicos y educativos*. Madrid: Síntesis.

Moreno R., Martínez, R., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, 16, 490-497.

Muñiz, J. (2003). *Teoría Clásica de los Tests*. Madrid: Pirámide.