Papeles del Psicólogo, 2010. Vol. 31(1), pp. 67-74
http://www.cop.es/papeles

FIABILIDAD Y VALIDEZ
RELIABILITY AND VALIDITY

Gerardo Prieto y Ana R. Delgado
Universidad de Salamanca

En este capítulo se describen conceptualmente las propiedades psicométricas de fiabilidad y validez y los procedimientos para evaluarlas. El apartado dedicado a la fiabilidad o precisión de las puntuaciones de las pruebas describe los distintos modelos, procedimientos empíricos e índices estadísticos para cuantificarla. En cuanto a la validez, la propiedad psicométrica más importante y la que ha experimentado mayores transformaciones a lo largo de la historia de la Psicometría, se resumen las principales concepciones y los debates en torno a la misma.
Se previene al lector de dos frecuentes malentendidos: en primer lugar, considerar que la fiabilidad y la validez son características de los tests cuando corresponden a propiedades de las interpretaciones, inferencias o usos específicos de las medidas que esos tests proporcionan; en segundo lugar, tratar la fiabilidad y la validez como propiedades que se poseen o no en lugar de entenderlas como una cuestión de grado.
Palabras clave: *Fiabilidad, Psicometría, Tests, Validez.*

The psychometric properties of reliability and validity and the procedures used to assess them are conceptually described in this chapter. The part devoted to the reliability, or test score accuracy, is focused in the models, procedures and statistical indicators most usually employed. As to validity, the most important psychometric property, and the one whose conception has changed the most, we summarize its history in testing contexts.
The reader is prevented that reliability and validity are not, as usually thought, properties of the testing instruments but of the particular inferences made from the scores. Another common error is considering reliability and validity, not as questions of degree, but as absolute properties.
Key words: *Reliability, Psychometrics, Testing, Validity.*

Los psicólogos utilizan diversos procedimientos estandarizados para obtener muestras de la conducta de las personas. Estos recursos, genéricamente denominados *tests*, incluyen un procedimiento de puntuación que permite obtener medidas que pueden ser usadas con distintos propósitos: estimar el nivel de la personas en un constructo (ansiedad, calidad de vida, visualización espacial…), evaluar la competencia tras un periodo de aprendizaje, clasificar a los pacientes en categorías diagnósticas o seleccionar a los aspirantes más aptos para un puesto de trabajo. La legitimidad y eficiencia de estas prácticas depende de su fiabilidad y validez.

En este capítulo se describen, de forma conceptual, estas dos características psicométricas y los procedimientos más frecuentes para evaluarlas. De entrada, hay que prevenir al lector de dos frecuentes malentendidos. El primero consiste en considerar que la fiabilidad y la validez son características de los tests. Por el contrario, corresponden a propiedades de las interpretaciones, inferencias o usos específicos de las medidas que esos tests proporcionan. El segundo se refiere a la consideración de que la fiabilidad y la validez se poseen o no, en lugar de entenderlas como una cuestión de grado (AERA, APA y NCME, 1999).

Correspondencia: Gerardo Prieto. *Facultad de Psicología. Universidad de Salamanca. Avda. De la Merced 109-131, 37005 Salamanca. España. Email: gprieto@usal.es*

FIABILIDAD

La fiabilidad se concibe como la consistencia o estabilidad de las medidas cuando el proceso de medición se repite. Por ejemplo, si las lecturas del peso de una cesta de manzanas varían mucho en sucesivas mediciones efectuadas en las mismas condiciones, se considerará que las medidas son inestables, inconsistentes y poco fiables. La carencia de precisión podría tener consecuencias indeseables en el coste de ese producto en una ocasión determinada. De esta concepción se sigue que de la variabilidad de las puntuaciones obtenidas en repeticiones de la medición puede obtenerse un indicador de la fiabilidad, consistencia o precisión de las medidas. Si la variabilidad de las medidas del objeto es grande, se considerará que los valores son imprecisos y, en consecuencia, poco fiables. De manera semejante, si una persona contestase a un test repetidamente *en las mismas condiciones*, de la variabilidad de las puntuaciones podría obtenerse un indicador de su grado de fiabilidad. La imposibilidad de lograr que las medidas se lleven a cabo *exactamente* en las mismas condiciones es

uno de los problemas de las medición psicológica y educativa. El nivel de atención y de motivación de una persona puede variar al contestar repetidamente a la misma prueba, la dificultad de dos tests pretendidamente iguales construidos para medir el mismo constructo puede ser desigual, las muestras de examinadores que califican un examen de selectividad pueden diferir en el grado de severidad, etc. Por tanto, el esfuerzo de los evaluadores ha de centrarse en estandarizar el procedimiento de medición para minimizar la influencia de aquellas variables extrañas que pueden producir inconsistencias no deseadas. La estandarización del procedimiento implica obtener las medidas en todas las ocasiones en condiciones muy semejantes: con el mismo tiempo de ejecución, las mismas instrucciones, similares ejemplos de práctica, tareas de contenido y dificultad equivalentes, similares criterios de calificación de los evaluadores de exámenes, etc.

El estudio de la fiabilidad parte de la idea de que la puntuación observada en una prueba es un valor concreto de una variable aleatoria consistente en todas las posibles puntuaciones que podrían haber sido obtenidas por una persona en repeticiones del proceso de medida en condiciones semejantes (Haertel, 2006). Obviamente, no es posible repetir la medición un número muy grande de veces a los mismos participantes. Por tanto, la distribución de las puntuaciones es hipotética y sus propiedades deben ser estimadas indirectamente. La media de esa distribución, que reflejaría el nivel de una persona en el atributo de interés, es denominada *puntuación verdadera* en la Teoría Clásica de los Tests (TCT). La TCT es un conjunto articulado de procedimientos psicométricos desarrollados fundamentalmente en la primera mitad del siglo pasado, que se ha utilizado extensivamente para la construcción, análisis y aplicación de los tests psicológicos y educativos. Aunque la TCT surgió en el contexto de la medición de las aptitudes humanas, sus propuestas se extienden a otras áreas. Se asume que la puntuación verdadera de una persona no cambia entre ocasiones, por lo que la variabilidad de las puntuaciones observadas se debe a la influencia de un *error de medida* aleatorio, no sistemático (producido por causas desconocidas e incontrolables en esa situación). La cantidad de error en cada caso sería la diferencia entre una puntuación observada y la puntuación verdadera. La desviación típica de los errores, denominada *error típico de medida* (ETM), indica la precisión de las puntuaciones de una persona, es decir, su variabilidad en torno a la puntuación verdadera. El ETM refleja el error que puede esperarse en una puntuación observada. Por ejemplo, si el error típico de medida del peso de un objeto fuese de dos gramos, se puede aventurar que el peso observado diferirá del peso verdadero en más de dos gramos solo la tercera parte de las veces. Aunque la TCT permite estimar el ETM para personas situadas en distintos rangos de la variable (denominados errores típicos de medida *condicionales*), suele emplearse un único valor aplicable de forma general a todas las puntuaciones de las personas de una población. Obviamente, la valoración del ETM depende de la magnitud de los objetos que se están midiendo: dos gramos es un error despreciable si se pesan objetos muy pesados como sacos de cereales, pero es un error notable si se pesan objetos más livianos como los diamantes. Es decir, el valor del ETM está en las mismas unidades que los objetos medidos y carece de un límite superior estandarizado que facilite su valoración. Por ello, se ha propuesto un índice estandarizado de consistencia o precisión denominado *coeficiente de fiabilidad* que puede oscilar entre 0 y 1. De la TCT se deriva que este coeficiente es el cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones observadas en una población de personas. En consecuencia, indica la proporción de la variabilidad de las puntuaciones observadas que *no* puede atribuirse al error de medida; por ejemplo, si el coeficiente de fiabilidad es de 0,80, se considera que el 20% de la variabilidad observada es espuria.

Para estimar empíricamente los estadísticos de fiabilidad (ETM y coeficiente de fiabilidad) se emplean diversos diseños de recogida de datos que reflejan distintas repeticiones del proceso de medida. Los más conocidos se denominan *test-retest* (aplicación de un test a una muestra de personas en dos ocasiones entre las que el atributo se mantiene estable), *formas paralelas* (aplicación a una muestra de personas en la misma ocasión o en distintas ocasiones de dos versiones del test equivalentes en contenido, dificultad, etc), *consistencia entre las partes de una prueba* (división del test en dos subconjuntos equivalentes de ítems o estimación a partir de las covarianzas entre los ítems de la prueba) y *consistencia de las puntuaciones de distintos calificadores* (evaluación de una muestra de conducta por calificadores independientes). La estimación del coeficiente de fiabilidad a partir de estos diseños suele basarse en la correlación entre las puntuaciones observadas obtenidas en las distintas formas de replicación. Existe una extensa bibliografía para obtener una información detallada de estos procedimientos y de los conceptos y desarrollos de la TCT. Excelentes exposiciones pueden encontrarse en este

volumen (Muñiz, 2010) y en los textos de Gulliksen (1950), Martínez-Arias, Hernández-Lloreda y Hernández-Lloreda (2006) y Muñiz (1998).

Además de la TCT, se emplean otros enfoques para cuantificar la fiabilidad de las puntuaciones de los tests: la Teoría de la Generalizabilidad (TG) y la Teoría de Respuesta al Ítem (TRI).

La TCT permite cuantificar solamente dos componentes de la varianza de las puntuaciones observadas: la varianza verdadera y la varianza de error. La TG, concebida como una extensión de la TCT, trata de especificar la contribución a la varianza observada de un número mayor de facetas: la variabilidad entre las personas, las ocasiones en que se mide, las diferentes formas del instrumento, los diferentes calificadores y las interacciones entre los componentes. La estimación de estas influencias se lleva a cabo mediante el análisis de varianza. Los componentes distintos a las diferencias entre personas (formas del test, calificadores, ocasiones, etc) se interpretan como fuentes del error de las medidas, sirviendo como evidencia de las posibles causas del error y permitiendo mejorar los procedimientos de medición. Este modelo es especialmente útil para evaluar la fiabilidad de las calificaciones otorgadas por evaluadores a los productos obtenidos en pruebas o exámenes *abiertos* (los examinados no están constreñidos por un formato cerrado, tal como los de las pruebas de elección múltiple, para emitir sus respuestas). Un tratamiento más exhaustivo puede encontrarse en los textos de Brennan (2001) y en este volumen (Martínez-Arias, 2010).

La TRI es un conjunto de modelos de medida dirigidos a estimar estadísticamente los parámetros de las personas y los ítems en un continuo latente a partir de las respuestas observables. En todos los procedimientos de estimación estadística de parámetros, se cuantifica la cantidad de error de la estimación a partir del error típico (un índice de la variabilidad de los estimadores del parámetro). Cuanto mayor sea el error típico, menor será la precisión de la estimación y mayor será la incertidumbre sobre el valor del parámetro. De forma similar, en los modelos de la TRI la incertidumbre sobre la localización de una persona o un ítem en la variable latente se cuantifica a partir del *error típico de estimación* (ETE) de la persona o del ítem. Este estadístico se diferencia del error típico de medida de las personas correspondiente a la TCT. Como ya se ha expuesto, el ETM es una medida *global* del error, un único valor aplicable de forma general a todas las puntuaciones de las personas de una población, que suele subestimar o sobrestimar el grado de error que afecta a las puntuaciones localizadas en distintos niveles de la variable. Por el contrario, el ETE varía a lo largo de la variable. Por tanto, puede ser considerado una medida *individual* de la precisión, dado que indica la magnitud del error con la que se estiman los parámetros de las personas o los ítems situados en distintas posiciones del continuo latente. La función que describe cómo cambian los valores del ETE de las personas en los distintos niveles de la variable es especialmente útil para determinar los rangos en los que un test es más fiable y para determinar la fiabilidad de los puntos de corte empleados en la clasificación de personas en categorías diagnósticas o de rendimiento.

Puesto que el ETE permite cuantificar un intervalo para estimar el parámetro de una persona, será mayor la incertidumbre sobre su localización cuanto mayor sea el intervalo. Si se adopta la perspectiva opuesta, es decir, de cuánta *certidumbre* se dispone sobre la localización de la persona, entonces se cuantifica la denominada función de información que es análoga al recíproco de la varianza de error condicional de la TCT. La función de información del test indica en qué medida éste permite diferenciar entre las personas en los distintos niveles del atributo. Véase una exposición más detallada en de Ayala (2009).

Terminaremos este apartado con algunas consideraciones prácticas acerca de la interpretación y el uso de los estadísticos de fiabilidad, comenzando por responder a una de las preguntas más frecuentes de los usuarios de las pruebas: ¿qué grado de fiabilidad deben tener las puntuaciones para que su uso sea aceptable? Sin duda, la magnitud requerida depende de las consecuencias derivadas del uso de las puntuaciones. Cuando las puntuaciones vayan a emplearse para tomar decisiones que impliquen consecuencias relevantes para las personas (p. ej., aceptación o rechazo en una selección de personal), el coeficiente de fiabilidad debería ser muy alto (al menos de 0,90). Sin embargo, si se trata de describir las diferencias individuales a nivel de grupo, bastaría con alcanzar valores más modestos (al menos 0,70). No obstante, estas convenciones deben seguirse con cautela: si la evaluación de la fiabilidad se ha llevado a cabo mediante los procedimientos derivados de la TCT, los resultados no habrán de ser necesariamente intercambiables, puesto que los diferentes diseños de recogida de datos antes mencionados (test-retest, formas paralelas, consistencia interna, etc) aprecian distintas fuentes de error: inestabilidad de las medidas, falta de equivalencia de las pruebas, heterogeneidad de los ítems, escasez de concordancia de los evaluadores, etc. Por tanto, es aconsejable disponer de estimaciones de la fiabilidad a partir

de distintos diseños para lograr una mejor comprensión del error que afecta a las puntuaciones (Prieto y Muñiz, 2000). Además, los estadísticos de fiabilidad varían entre poblaciones y están afectados por otras condiciones como la longitud de la prueba y la variabilidad de las muestras de personas. En consecuencia, se ha de evitar el error de considerar que la estimación de la fiabilidad procedente de un único estudio refleja la verdadera y única fiabilidad de la prueba. Los constructores y los usuarios de las pruebas deben informar detalladamente de los métodos de cuantificación, de las características de las muestras y de las condiciones en las que se han obtenido los datos (AERA, APA y NCME, 1999). Como hemos indicado anteriormente, el error típico de medida está expresado en las mismas unidades que las puntuaciones de la prueba. Por ello, es difícil establecer comparaciones entre la fiabilidad de las puntuaciones de distintos tests en base a este estadístico. Por el contrario, la magnitud del coeficiente de fiabilidad oscila siempre entre unos límites estandarizados (0 y 1), por lo que es muy útil para elegir el test más fiable entre los potencialmente utilizables para una aplicación específica. Sin embargo, el error típico de medida aporta más información para describir la precisión de las puntuaciones.

En ocasiones, se utilizan las puntuaciones de los tests, no simplemente para estimar la posición de una persona en la población de interés (denominada interpretación *relativa*), sino para asignarla a una categoría diagnóstica o de rendimiento (patológica/normal, apto/no apto, aceptado/excluido, etc). Para realizar este tipo *absoluto* de interpretaciones, se suelen emplear puntos de corte que guían la clasificación. Puesto que la fiabilidad de las puntuaciones no suele ser la misma en todos los niveles de la variable, conviene conocer el grado de error en las cercanías del punto de corte, dado que si es alto será elevado el número de falsos positivos y negativos en la clasificación. En este caso, es aconsejable emplear la función de error de estimación o de información derivada de los modelos de la TRI.

Terminaremos este apartado analizando la relación entre la fiabilidad y la validez de las puntuaciones, la propiedad que se describe en el siguiente apartado. En la actualidad se considera que la validez, definida como el grado en que las interpretaciones y los usos que se hacen de las puntuaciones están justificados científicamente, es la propiedad psicométrica más importante. Obviamente, la utilidad de unas puntuaciones escasamente fiables para tales fines estará seriamente comprometida. De ahí que se considere la fiabilidad como condición necesaria de la validez. Sin embargo, no será una condición suficiente si las puntuaciones verdaderas, aunque se estimen de manera muy precisa, no resultan apropiadas para conseguir el objetivo de la medida (representar un constructo, predecir un criterio de interés, etc). Es útil tener presente que la fiabilidad es una cuestión relativa a la calidad de los datos, mientras que la validez se refiere a la calidad de la inferencia (Zumbo, 2007).

VALIDEZ

El concepto de validez ha experimentado transformaciones importantes durante el último siglo, provocadas por los diversos objetivos a los que se han destinado los tests. De acuerdo con Kane (2006), entre 1920 y 1950 el uso principal de las pruebas consistió en predecir alguna variable de interés denominada *criterio* (por ejemplo, el rendimiento laboral o académico). En la actualidad este enfoque sigue siendo de suma importancia cuando se emplean las pruebas para seleccionar a los candidatos más aptos para un empleo, en los programas de admisión, en la adscripción de pacientes a tratamientos, etc. En estos casos, la evaluación de la utilidad de la prueba suele cuantificarse mediante la correlación entre sus puntuaciones y las de alguna medida del criterio (*coeficiente de validez*). Sin embargo, el éxito de este tipo de justificación depende de la calidad de la medida del criterio, especialmente de su representatividad (por ejemplo, ¿los indicadores para medir el criterio son suficientes y representativos del puesto de trabajo a desempeñar?). De ahí que el énfasis se desplazase a la justificación de que la puntuación en el criterio procedía de una muestra de indicadores que representase de forma apropiada el dominio o *contenido* a medir (la totalidad de los indicadores posibles). Por tanto, esta fase inicial de desarrollo del concepto terminó con la propuesta de dos vías regias para establecer la validez de las pruebas: la validación de criterio (la correlación entre las puntuaciones del test y las puntuaciones en el criterio) y la validación de contenido (la justificación de que los ítems para medir el criterio son una muestra representativa del contenido a evaluar).

La validación de contenido se extendió desde el análisis del criterio al de la validez de los tests predictores: una prueba no puede considerarse válida si los ítems que la componen no muestrean adecuadamente el contenido a evaluar. La validación de contenido es un enfoque especialmente fértil cuando las facetas del dominio a medir pueden identificarse y definirse claramente. Es éste el caso de los tests dirigidos a evaluar el rendimiento académico que puede especificarse en función de los objetivos de la instrucción (conceptos y ha-

bilidades que un alumno ha de poseer). La metodología de validación descansa fundamentalmente en la evaluación de expertos acerca de la pertinencia y la suficiencia de los ítems, así como de la adecuación de otras características de la prueba como las instrucciones, el tiempo de ejecución, etc. Sin embargo, especificar con precisión el contenido de las manifestaciones de constructos como la extraversión, la memoria de trabajo o la motivación de logro es una tarea más difícil. De ahí que tanto la validación de contenido como la de criterio se considerasen insuficientes para justificar el uso de pruebas dirigidas a evaluar aptitudes cognitivas o atributos de la personalidad. Esta insatisfacción se concretó en el influyente artículo de Cronbach y Meehl (1955) en el que se propone la validación de *constructo* como el modo principal de validación. Cronbach (1971) puntualizó que en un test para medir un rasgo de personalidad no hay únicamente un criterio relevante que predecir, ni un contenido que muestrear. Se dispone, por el contrario, de una teoría acerca del rasgo y de sus relaciones con otros constructos y variables. 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. En consecuencia, la validación de constructo puede concebirse como un caso particular de la contrastación de las teorías científicas mediante el método hipotético-deductivo. Aunque el usuario no sea, en general, consciente de ello, las técnicas de medida implican teorías (que se suponen suficientemente corroboradas en el momento de usarlas para contrastar hipótesis científicas o prácticas), por lo que deben venir avaladas ellas mismas por teorías cuyo grado de sofisticación dependerá del momento en que se encuentre el programa de investigación en el que han surgido (Delgado y Prieto, 1997). Dado que una teoría postula una red de relaciones entre constructos y atributos observables, no podremos asumir que las puntuaciones son válidas si la teoría es formalmente incorrecta, las predicciones derivadas de la teoría no se cumplen en los datos empíricos o se han violados otros supuestos auxiliares. Así, desde finales del siglo pasado se ha impuesto 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 (Messick, 1989). El marco de validación se define a partir de teorías en las que se especifican el significado del constructo a evaluar, sus relaciones con otros constructos, sus manifestaciones y sus potenciales aplicaciones e interpretaciones. Además de las pruebas necesarias para justificar una adecuada representación del constructo, Messick incluyó en el marco de validación la justificación de las *consecuencias* del uso de los tests (las implicaciones individuales y sociales). Como se comentará más adelante, la inclusión de la denominada *validación de las consecuencias* es aún objeto de debate. Este breve resumen de la historia del concepto de validez, de la que hemos mencionado algunos hitos importantes, permite comprender los conceptos actuales de validez y validación, de los que destacaremos a continuación sus principales características.

En la actualidad se considera que la *validez* se refiere al grado en que la evidencia empírica y la teoría apoyan la interpretación de las puntuaciones de los tests relacionada con un uso específico (AERA, APA y NCME, 1999). La *validación* es un proceso de acumulación de pruebas para apoyar la interpretación y el uso de las puntuaciones. Por tanto, el objeto de la validación no es el test, sino la interpretación de sus puntuaciones en relación con un objetivo o uso concreto. El proceso de validación se concibe como un *argumento* que parte de una definición explícita de las interpretaciones que se proponen, de su fundamentación teórica, de las predicciones derivadas y de los datos que justificarían científicamente su pertinencia. Dado que las predicciones suelen ser múltiples, una única prueba no puede sustentar un juicio favorable sobre la validez de las interpretaciones propuestas. Son necesarias pruebas múltiples y convergentes obtenidas en diferentes estudios. Por ello, se considera que la validación es un proceso dinámico y abierto. Obviamente, los usos y las interpretaciones relacionadas pueden ser muy variados. Por ello, las fuentes de validación son múltiples y su importancia varía en función de los objetivos. Los *Standards for educational and psychological testing* (AERA, APA y NCME, 1999) se refieren a las más importantes: el contenido del test, los procesos de respuesta, la estructura interna de la prueba, las relaciones con otras variables y las consecuencias derivadas del uso para el que se proponen. Antes de resumir estos enfoques metodológicos, hemos de puntualizar que reflejan distintas facetas de la validez que las engloba como un único concepto integrador. Por tanto, no es riguroso utilizar términos, como *validez predictiva, validez de contenido, factorial*, etc, que inducirían a considerar distintos tipos de validez.

Validación del contenido del test

Los tests están compuestos por un conjunto de ítems destinados a obtener una puntuación que represente el nivel de una persona en un constructo (extraversión,

competencia en matemáticas, etc). Difícilmente se podrá justificar la calidad de las medidas si los ítems no representan de forma suficiente las diferentes facetas de las manifestaciones del constructo. Si eso sucede, el constructo estará *infrarrepresentado* y, en consecuencia, las puntuaciones no alcanzarán el grado de validez requerido. Asimismo, la evidencia de que las respuestas a los ítems están influidas por variables ajenas al constructo de interés constituye una de las principales amenazas a la validez produciendo la denominada *varianza irrelevante al constructo*. También son objeto de la validez de contenido las instrucciones, los ejemplos de práctica, el material de la prueba, el tiempo de ejecución, etc. La consulta a expertos es la vía más usual para apreciar la calidad del contenido, especialmente en ámbitos educativos, aunque cada vez son más empleados los métodos cualitativos basados en la observación directa, las entrevistas o el análisis de archivos. Los procedimientos estandarizados de consulta facilitan la obtención de datos cuantitativos indicativos del porcentaje de ítems de calidad, el porcentaje de las facetas del dominio suficientemente evaluadas, el porcentaje de jueces que han valorado positivamente la calidad de los materiales, la concordancia entre los expertos, etc. Un tratamiento exhaustivo de la validación del contenido puede encontrarse en Sireci (1998).

Análisis de los procesos de respuesta

Debido a la influencia de la ciencia cognitiva, la validación de los tests de inteligencia, aptitudes y rendimiento debe incluir el análisis de los procesos, las estrategias de resolución de problemas y las representaciones mentales que emplean los participantes para resolver los ítems. Se obtendrá evidencia de validez cuando los procesos utilizados se ajustan a los que se postulan en las teorías relativas al constructo medido. La metodología de estudio es muy diversa: entrevistas a los examinados para que describan cómo resuelven las tareas, análisis de los movimientos oculares o tiempos de respuesta, etc. Cuando las teorías acerca del constructo han superado las etapas meramente exploratorias, se pueden construir los tests a partir de un *diseño cognitivo* que especifica ciertos subconjuntos de ítems para suscitar determinados procesos latentes. Las respuestas a los ítems permiten estimar, mediante modelos complejos de la TRI, los parámetros de la persona en los distintos componentes cognitivos de la tarea e identificar *clases* de personas que emplean distintas estrategias de procesamiento. En este enfoque se basan las tendencias más avanzadas del diagnóstico cognitivo (Yang y Embretson, 2007).

Análisis de la estructura interna del test

Algunos tests proporcionan una medida de un solo constructo, otros evalúan varios constructos incluyendo una subescala para cada uno de ellos. El análisis de la estructura interna persigue verificar empíricamente si los ítems se ajustan a la dimensionalidad prevista por el constructor de la prueba. Cuando un test construido inicialmente para evaluar a las personas de una población específica se pretende adaptar a una población diferente (de otra cultura, por ejemplo), es obligado analizar si la estructura interna de la prueba se mantiene invariante. En caso contrario, el significado de las puntuaciones diferirá entre ambas poblaciones. El análisis de la estructura interna del test suele llevarse a cabo con ayuda de los modelos de análisis factorial que se describen en detalle en el artículo de Ferrando y Anguiano (2010) de este monográfico.

Entre los métodos para evaluar la unidimensionalidad de la prueba, ocupa un lugar importante el análisis del *funcionamiento diferencial de los ítems* (DIF). Se podrá aseverar que un test tiene una validez similar en grupos de distinto sexo, cultura, lengua materna, etc., si sus ítems no presentan DIF, como puede leerse en el artículo de Gómez-Benito, Hidalgo y Guilera (2010).

Asociación de las puntuaciones con otras variables

Las relaciones de las puntuaciones del test con otras variables externas a la prueba constituyen una importante fuente de validación. Cuando se emplean las puntuaciones para seleccionar los candidatos más aptos para un empleo, en los programas de admisión, en la adscripción de pacientes a tratamientos, etc, la justificación se basa en su utilidad para predecir un criterio externo. El criterio es una medida de la variable de interés: rendimiento laboral, presencia o ausencia de un trastorno neuropsicológico, calificaciones académicas, etc. La utilidad de la prueba se suele cuantificar mediante la correlación entre sus puntuaciones y las de alguna medida del criterio (*coeficiente de validez*), o mediante otros procedimientos: diferencia en las puntuaciones entre grupos de distinto nivel en el criterio, grado de acuerdo en las clasificaciones en categorías diagnósticas realizadas mediante el test y expertos, etc. La elección de un criterio fiable y válido (suficiente, objetivo y representativo de la conducta de interés) es el punto crítico que determina la bondad del proceso de validación. En función del momento temporal en el que se evalúa el criterio, se distinguen distintos tipos de recogida de datos: *retrospectiva* (el criterio se ha obtenido antes de administrar el test, por ejemplo en base a un diagnóstico clínico anterior), *concurrente* (las puntuaciones del test y del criterio se obtienen en la misma sesión) y *predictiva* (el criterio se mi-

de en un momento posterior). Los resultados entre estos procedimientos pueden diferir: se preferirá el más adecuado al uso que se pretende (por ejemplo, el enfoque predictivo es más apropiado al pronóstico de un rendimiento laboral futuro). De suma importancia es analizar si la utilidad predictiva o diagnóstica se mantiene invariante en distintos grupos de personas. La cuestión de la variabilidad de los resultados en distintos grupos, distintos estudios, diferentes medidas del criterio, etc. afecta a la generalización de la validez de la prueba. El meta-análisis (véase el artículo de Sánchez-Meca y Botella, 2010) permite indagar cómo varían las correlaciones entre el test y el criterio en función de distintas facetas de los estudios.

Cuando las puntuaciones de los tests se usan para estimar el nivel de las personas en un constructo, sus correlaciones con las de otros tests que miden el mismo u otros constructos son de una relevancia especial. Se espera que la asociación entre pruebas que midan el mismo constructo, sean mayores (*validación convergente*) que entre tests que miden constructos diferentes (*validación discriminante*). Para obtener evidencia empírica, se emplean técnicas como el análisis factorial o la matriz multirrasgo-multimétodo (Campbell y Fiske, 1959) en la que se resumen las correlaciones de un test con *marcadores* (tests de validez comprobada) que miden varios constructos a través de distintos métodos.

Validación de las consecuencias del uso de los tests

La última versión de los *Standards for educational and psycological testing* (AERA, APA y NCME, 1999) plantea la previsión de las posibles consecuencias del uso de los tests como parte del proceso de validación. Desde esta perspectiva, el análisis y justificación de las consecuencias ocupan un lugar preponderante cuando los tests vayan a emplearse para tomar decisiones críticas para personas e instituciones: selección, contratación, graduación, promoción profesional, evaluación de programas, etc. La literatura psicométrica denomina estos usos como de *alto riesgo*. Estas prácticas no son ajenas al contexto español: selección de los candidatos a piloto, al ejército profesional y los cuerpos de seguridad, oposiciones para ingresar en diversas instituciones y empresas, exámenes universitarios, pruebas de selectividad, evaluación del profesorado universitario, evaluación del grado de dependencia, obtención del permiso de armas y del carnet de conducir, etc. En estos casos, la pertinencia del uso no se limita a la comprobación de que las puntuaciones representan adecuadamente los constructos y a la justificación teórica de la red nomológica que vincula los constructos con los criterios de interés. Las aplicaciones de alto riesgo tienen efectos colaterales de carácter personal y social. Citemos como ejemplo de los primeros el efecto en la validez de las puntuaciones del entrenamiento y aprendizaje de los tests que suelen seguir muchas de las personas que se presentan a programas de selección. ¿Hasta qué punto son sensibles las pruebas a este tipo de manipulación? Existen otros efectos de carácter institucional tales como la peculiaridad del uso de los tests en un contexto social. Piénsese en el fraude social relacionado con el uso de las pruebas psicotécnicas que se emplean en nuestro país para otorgar el permiso de armas o el de conducir. Si pensamos en las consecuencias, ¿podríamos decir que ejercen su función? Está claro que si la validez se refiere al grado en que la teoría y la evidencia empírica apoyan la interpretación de las puntuaciones de los tests en relación con un uso específico, las consecuencias no pueden ser ajenas al proceso de validación.

Aunque parece existir un cierto consenso sobre esta cuestión, también existen voces discordantes. Por ejemplo, Borsboom y Mellenberg (2007) consideran que el concepto de validez debe limitarse a un ámbito más restringido que el de la amplia definición incluida en las propuestas de Messick (1989) y en los actuales *Standards*. A su juicio, la validación debe limitarse a contrastar si existe una relación causal entre el constructo y las puntuaciones del test; las interpretaciones de las puntuaciones en contextos aplicados (selección de personal, acreditación, etc) y el impacto social del uso de las pruebas serían ajenas, *stricto sensu*, al ámbito de la validez. Si bien esta postura simplificadora parece libre de problemas, definir la validez de constructo como la validez de la inferencia causal implica identificarla con la validez interna de la evidencia a favor del constructo (para una versión actualizada de los distintos tipos de validez en los diseños experimentales véase Shadish, Cook y Campbell, 2002). Esta identificación podría, tal vez, justificarse en programas de investigación básica ya avanzados, pero imposibilitaría en la práctica la mayor parte de las aplicaciones psicológicas, y esto sin tener en cuenta los conocidos problemas del concepto de causación. De ahí que el pragmatismo nos lleve a preferir una postura más flexible, la que considera que los procedimientos de validación han de servir para apoyar la inferencia a la mejor explicación posible, incluyendo la evidencia aportada por los diversos métodos cualitativos y cuantitativos a disposición de los psicómetras en cada momento (Zumbo, 2007). Si se considera que la validación es un proceso abierto en el tiempo, la validez es necesariamente una cuestión de grado, como señalan los *Standards*, algo que, por otra parte, es común a los distintos conceptos de validez empleados por los epistemólogos.

El debate sobre la inclusión de las consecuencias en el concepto de validez no es un tecnicismo que preocupe solo a los sesudos teóricos de la psicometría. Tomar partido por la inclusión conlleva responsabilidades: ¿pueden y deben los constructores de las pruebas aventurar las consecuencias deseables e indeseables de su uso? ¿qué repertorio metodológico usar para ello? ¿en qué instancia recae el análisis y la justificación de las consecuencias? Estas y otras cuestiones relacionadas seguirán alimentando el debate y la generación de propuestas: una excelente revisión sobre la validación de las consecuencias puede consultarse en Padilla, Gómez, Hidalgo y Muñiz (2007).

Para terminar, un comentario terminológico: acorde con la trayectoria del uso de los tests en contextos anglosajones, *validation* tiene en inglés un significado legal: "declarar legalmente válido". Por el contrario, en nuestra lengua, el término validación tiene dos significados: "acción y efecto de validar", que comparte con el idioma inglés, y "firmeza, fuerza, seguridad o subsistencia de algún acto". Aunque solemos referirnos a la primera acepción, la más aséptica, es la segunda la que más se acerca al objetivo que persigue la investigación psicológica en su variante psicométrica.

REFERENCIAS

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Borsboom, D. y Mellenberg, G.J. (2007). Test Validity in Cognitive Assessment. En J.P. Leighton y M.J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 85-115). Cambridge: Cambridge University Press.

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

de Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

Campbell, D.T. y Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cronbach, L. J. (1971). Test validation. En R.L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J. y Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Delgado, A.R. y Prieto, G. (1997). *Introducción a los métodos de investigación de la psicología.* Madrid: Pirámide.

Gómez-Benito, J., Hidalgo, M.D. y Guilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo*, *31(1)*, 75-84.

Gulliksen, H. (1950). *Theory of mental tests*. New York, Wiley.

Haertel, E. H. (2006). Reliability. En R.L. Brennan (Ed.), *Educational Measurement* (pp. 65-110). Wesport, CT: American Council on Education and Praeger Publishers.

Kane, M.T. (2006). Validation. En R.L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Wesport, CT: American Council on Education and Praeger Publishers.

Martínez-Arias, M.R. (2010), Evaluación del desempeño. *Papeles del Psicólogo*, *31(1)*, 85-96.

Martinez-Arias, M.R., Hernández-Lloreda, M.J. y Hernández-Lloreda, M.V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Messick, S. (1989). Validity. En R.L. Linn (Ed.), *Educational measurement* (pp- 13-103). New York: American Council on Education.

Muñiz, J. (1998). *Teoría Clásica de los Tests*. Madrid: Pirámide.

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo*, *31(1),* 57-66.

Padilla, J.L., Gómez, J., Hidalgo, M.D. y Muñiz, J. (2007). Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los tests. *Psicothema, 19*, 173-178.

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77,* 65-71.

Sánchez-Meca, J. y Botella, J. (2010). Revisiones sistemáticas y meta-análisis: herramientas para la práctica profesional. *Papeles del Psicólogo*, *31(1)*, 7-17.

Shadish, W.R., Cook, T.D., y Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.

Sireci, S.G. (1998). The construct of content validity. En Zumbo, B.D. (Ed.), *Validity Theory and the Methods Used in Validation: Perspectives From the Social and Behavioral Sciences* (pp. 83-117). Kluwer Academic Press, The Netherlands.

Yang, X. y Embretson, S.E. (2007). Construct Validity and Cognitive Diagnostic Assessment. En J.P. Leighton y M.J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 119-145). Cambridge: Cambridge University Press.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics,* (pp. 45-79). Elsevier Science B.V.: The Netherlands.