Psicothema 2003. Vol. 15, nº 2, pp. 315-321
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Sobre la validez de los tests

Paula Elosua Oliden
Universidad del País Vasco

Han transcurrido más de dos años desde la publicación de los últimos estándares sobre el uso de tests psicológicos y educativos (AERA, APA y NCME, 1999). Su contenido es el referente más claro y ortodoxo para la evaluación, construcción y utilización de los tests. Desde la versión anterior, publicada en 1985, la aportación más relevante se centra en la necesidad de garantizar un uso correcto de los tests. Es un nuevo matiz que otorga al usuario responsabilidades hasta ahora no consideradas. Como consecuencia irrumpen nuevas fuentes de evidencia en el análisis de la validez: el funcionamiento diferencial de los ítems y la validez consecuencial. El objetivo de este trabajo es ofrecer una panorámica general sobre la validez centrada en los últimos estándares, que la definen como el aspecto más relevante tanto en el desarrollo como en la evaluación de los tests.

About test validity. More than two years have passed since the publication of the last standards for educational and psychological testing (AERA, APA y NCME, 1999). This publication is the best and foremost reference for test evaluation, construction and use. The most important contribution to the previous version published in 1985 is the emphasis in guaranteeing proper use of the tests, thus bestowing upon the user new responsibilities in the process. As a result of this, new sources of evidence for validity analysis such as differential item functioning and consequential validity have prouted strongly. The objective of this study is to provide a general overview on «validity» based on the latest standards, which is the most relevant feature for both test development and test evaluation.

La validez es el aspecto de la medición psicopedagógica vinculado con la comprobación y estudio del significado de las puntuaciones obtenidas por los tests. Acorde a una orientación marcadamente empírica, la psicología actual centra su estudio en el examen de las variables definidas en y por el test, y de sus relaciones con variables externas, observadas o latentes, con el objeto de sustentar las interpretaciones propuestas.

La evolución de su significado desde un origen pragmático y operacional, hasta la complejidad de la visión que hoy impera, refleja el carácter progresivo de la ciencia que la cobija. Se ha revestido de mil formas, acepciones o enfoques (convergente, discriminante, factorial, sustantiva, estructural, externa, de población, ecológica, temporal, de tarea (Messick, 1980)) bajo las cuales es posible delimitar *grosso modo*, tres etapas que han quedado impresas en la redacción de los estándares de 1974, 1985 y 1999:

- Una primera etapa *operacional* dominada por una visión pragmática que prima la validez extema («Un test es válido para aquello con lo que correlaciona»; Guilford, 1946; p. 429). Esta perspectiva diferencia entre 4 tipos de validez: contenido, predictiva, concurrente y de constructo.(APA; AERA; NCME, 1954), que las ediciones de 1966 y 1974 (APA, AERA, NCME, 1966; 1974) reducen a tres agrupando para ello la validez predictiva y concurrente en la validez referida al criterio. La visión tripartita admitida no se romperá oficialmente hasta la publicación de los estándares de 1985.
- Un segundo estadio *teórico* marcado por la importancia concedida a la teoría psicológica, en el que se adopta una visión integradora. Se impone el análisis de la validez de constructo como concepto unificador que abarca aspectos de contenido y de relaciones con otras variables. («..toda validación es validación de constructo»; Cronbach, 1984; p. 126). Supone el reconocimiento de la validez como proceso único de recogida de evidencias a través de estrategias de investigación diferentes relacionadas con el constructo, con el contenido o con el criterio.
- La fase actual, a la que podríamos denominar, *contextual*, en la que se amplía la acepción anterior y se delimita con el concepto de *uso propuesto*. Su objetivo sería dotar a los tests de avales tanto científicos como éticos. («Una visión integradora de la validez…debe distinguir dos facetas interconectadas del concepto unitario de validez. Una faceta es la fuente de justificación… La otra faceta es la función o resultado del test…»; Messick, 1989; p. 20). En esta nueva revisión no se encuentran referencias a distintas formas de validez. Se incorpora a la connotación teórica anterior un aspecto hasta entonces olvidado, el uso. Ya no es suficiente la justificación sustantiva de las puntuaciones, es necesario delimitar los fundamentos teóricos en un contexto externo, con relación al *propósito* o *interpretación propuesta*. Como consecuencia, dentro de los ámbitos de uso de un test (Tabla 1) habrán de especificarse las condiciones de la situación de medida, que entre otros aspectos, tendrán en cuenta la rele-

Fecha recepción: 28-5-02 • Fecha aceptación: 8-11-02
Correspondencia: Paula Elosua Oliden
Facultad de Psicología
Universidad del País Vasco
20009 San Sebastián (Spain)
E-mail: pspelolp@sc.ehu.es

vancia y utilidad de las puntuaciones para los fines propuestos.

Bajo esta postura descansa la aseveración de que interpretar un test es usarlo, y de que todos los usos incluyen una interpretación del test. De ahí que se confiera a la persona responsable de la administración del test un estatus privilegiado e irremplazable para el análisis del significado y relevancia de las puntuaciones. Aunque la descripción de las variables que influyen sobre éstas forma parte de la fase de construcción, el usuario habrá de reconocer los posibles factores contaminantes que operan en cada situación particular. De este modo, sobre éste recae una carga tanto ética como interpretativa. El constructor justificará teóricamente el uso, pero es el agente final el que habrá de valorar la adecuación del contexto a los requerimientos de validez.

La importancia concedida a las implicaciones derivadas de una contextualización práctica o uso está estrechamente ligada al concepto de *sesgo*. Es un término con connotaciones políticas, sociales, estadísticas y psicométricas, que comienza a cobrar relevancia en la década de los 20 debido a la controversia surgida en Estados Unidos acerca de la parcialidad de los tests respecto a determinados grupos (Jensen, 1980).

Desde un punto de vista estrictamente psicométrico el sesgo es un error sistemático originado por deficiencias en el test o en el modo en que éste es usado, que produce una distorsión en el significado de las puntuaciones y que adultera la interpretación propuesta. Sesgo y validez se convierten en aspectos afines. El sesgo siempre supondrá falta de validez, y la falta de validez puede ser el origen del sesgo. Para maximizar una y consecuentemente minimizar otra, el test habrá de incorporar una descripción detallada de cada uno de los ámbitos de uso propuestos, que servirá de marco conceptual básico para la recopilación de evidencias e interpretación de puntuaciones. El fin es comprobar que no existen ni infrarrepresentación del constructo ni varianza irrelevante para el mismo causadas por la intervención de variables ajenas tanto al marco teórico como a los objetivos propuestos. Con la inclusión de estos aspectos dentro del proceso de validación se adopta un punto de vista multidimensional sobre el origen del sesgo. Un instrumento de medida, o en este caso sería más correcto hablar del uso de un instrumento de medida, puede ser origen de sesgo si su estructura interna y distribución difieren entre grupos.

El objetivo de los estudios de validez sería por todo ello recoger las suficientes evidencias que pueden prestar una base científica a la interpretación de las puntuaciones en un uso concreto. Estas pueden provenir de diversas fuentes. La importancia otorgada a cada una de ellas dependerá de los objetivos del test, que serán en cada caso los que determinarán las más significativas. Los últimos estándares diferencian entre fuentes relacionadas con el contenido, el proceso de respuesta, la estructura interna, las relaciones con otras variables y las consecuencias del test. Podemos agruparlas en fuentes de evidencia internas y externas. Las primeras suponen un análisis individualizado de los ítems, mientras que las segundas analizan el test en conjunto.

Fuentes de evidencia internas

Contenido

El análisis del contenido aglutina dos tipos de estudios suplementarios. Unos encaminados a evaluar las relaciones entre el constructo y el contenido del test, y otros dirigidos a valorar los factores contextuales internos y externos que puedan añadir varianza no deseada.

El objetivo de los primeros es garantizar que la muestra de ítems que componen la prueba es además de relevante, representativa del constructo. Su análisis incluye tres aspectos, la definición del dominio, y el estudio de su representación y de su relevancia (Sireci, 1998). El primero se centra en la definición operacional del dominio del contenido, que tradicionalmente se sirve de una tabla bidimensional en la que se especifican las áreas de contenido y las áreas cognitivas que se pretenden evaluar. La representación y relevancia, por su parte, consisten en la evaluación de

Tabla 1
Ámbitos de uso de los tests (AERA, APA, NCME, 1999)

Ámbitos de uso	**Propósito**	
Evaluación psicológica	Diagnóstico Intervención Decisiones jurídicas Crecimiento personal Selección individual	
Evaluación educativa	Diagnóstico individual	Rendimiento y cambio en un dominio de contenido Carencias Planificación de intervenciones Inclusión en programas de apoyo Selección de candidatos Certificaciones
	Diagnóstico colectivo	Evaluación de programas educativos Evaluación de políticas o intervenciones educativas
Empleo y acreditación	Selección Promoción Ubicación Evaluación de aptitudes y competencias Evaluación de Programas	

cada uno de los ítems en función de la definición dada. La evidencia basada en el contenido, aunque en su mayoría cualitativa y sustentada en análisis lógicos, puede incluir, sobre todo en tests de rendimiento y referidos al criterio, índices empíricos de congruencia basados en pruebas inter-jueces o en técnicas de escalamiento uni- y multidimensional (Hambleton, 1980).

El estudio de los factores contextuales cubre un amplio abanico de condiciones que abarcan entre otras, el formato de los ítems, el tipo de tareas exigidas, y la evaluación de la propia situación de test. Dentro de esta última se incluirían las instrucciones para la administración y corrección de la prueba, la interacción entre examinador-examinado, la familiaridad con la situación, las diferencias de motivación o ansiedad o el tipo de material utilizado. El objetivo es evitar fuentes de dificultad irrelevantes o un uso sesgado del lenguaje para lo cual se aconseja evaluar las distintas acepciones o significados que un mismo término puede poseer para diferentes grupos y asegurar que la experiencia curricular de los sujetos sea la misma.

Proceso de respuesta

La influencia ejercida por la psicología cognitiva sobre la psicometría tradicional está obligando a reanalizar la medición por medio de tests para que fije su atención más que en la utilidad del constructo en su representación (Prieto y Delgado, 1999; Snow y Lohman, 1993). En la búsqueda de instrumentos de medida que se ajusten a un marco que no sea estrictamente estadístico, el estudio de los procesos cognitivos implicados en la resolución de los ítems es un importante foco de información.

La metodología descansa en los protocolos de respuesta, entrevistas, y en general procedimientos que permitan el análisis individualizado del par sujeto/ítem. Desde la teoría de respuesta al ítem se han propuesto diversos modelos, los componenciales, para acometer este fin. Son formulaciones que aúnan la representación formal y la psicológica, descomponiendo la dificultad de los ítems en parámetros representativos de sus componentes (Embretson, 1997). Se trata de instrumentos útiles para la constatación de modelos cognitivos que permiten además indagar posibles discrepancias entre grupos referidas al procesamiento de las respuestas.

Estructura interna

En el intervalo de 14 años transcurrido entre las dos últimas revisiones de los estándares, los aspectos relacionados con el análisis de la estructura interna son tal vez los que más literatura especializada han originado. Esta fuente de evidencia evalúa el grado en que las relaciones entre los ítems y los componentes del test conforman el constructo que se quiere medir y sobre el que se basarán las interpretaciones. Podría asimilarse al aspecto interno de la validez de constructo definida por Loevinger (1957), o a la representación del constructo apuntada por Embretson (1983). Según los últimos estándares, se centra en la evaluación de la dimensionalidad de la prueba, y del funcionamiento diferencial de los ítems.

Dimensionalidad

El estudio del número de factores, dimensiones o habilidades subyacentes a un conjunto determinado de variables es uno de los

Tabla 2
Fuentes de evidencia

Evidencia	Tipo	Método
INTERNA (ÍTEMS)	CONTENIDO	Definición del dominio. Representación y relevancia Situación de test (formato, administracion, puntuación)
	PROCESO DE RESPUESTA	Protocolos Entrevistas Modelos componenciales
	ESTRUCTURA INTERNA **Dimensionalidad**	Modelos de estructura latente Modelo Factor Común Modelo Respuesta ítem Paramétrico No-paramétrico
	F.D.I.	Invarianza observada Delta, chi-cuadrado, Mantel-Haenszel, Regresión logística, Log-Lineal, SIBTEST Invarianza Latente Modelo Respuesta al ítem Modelo Factor Común
EXTERNA (TEST)	RELACIONES **Convergente/discriminante**	Matriz multirrasgo/multimétodo Factorial Confirmatorio
	Test/criterio	Modelo lineal generalizado
	Generalización	Meta-análisis
	CONSECUENCIAS	

temas más recurrentes de la psicometría. Su objetivo es la determinación del mínimo número de estructuras necesario para explicar la máxima varianza observada. Se trataría de definir un modelo linealmente independiente y monótono a través de un número reducido de factores (Stout, 1990).

Las perspectivas que pueden adoptarse para la especificación dimensional pueden englobarse bajo el término genérico de modelos de rasgo latente (McDonald, 1999). Dentro de ellos situaríamos los modelos lineales derivados del modelo del factor común, y los no-lineales procedentes de los modelos de respuesta al ítem.

De entre todos ellos, el análisis factorial es el que ha gozado de mayor popularidad. Basado en el modelo lineal del factor común de Spearman, integra un conjunto de técnicas de análisis multivariadas cuya finalidad es resumir la información contenida en un conjunto de variables observadas por medio de un número reducido de variables hipotéticas, conocidas habitualmente como factores. El objetivo es reproducir las matrices de covarianzas o correlaciones entre variables observadas .

Sin embargo, la linealidad en las relaciones variable/factor que asume este modelo se viola en muchas de las situaciones analizadas en psicología. La relación entre una variable dicotómica y un factor, por ejemplo, nunca es lineal. Ante estos casos, y desde los modelos de respuesta al ítem se desarrollan los modelos multidimensionales compensatorios, que se han mostrado especialmente útiles en la determinación de la estructura interna de datos dicotómicos (Elosua y López, 2002; Hambleton y Rovinelli, 1986; Hattie, 1984). Son modelos no-lineales, logísticos o de ojiva, que permitiendo una doble parametrización (factorial, de respuesta al ítem), ejercen una función de nexo entre dos acercamientos que aunque aparentemente divergentes presentan grandes similitudes estructurales.

El estudio de la dimensionalidad no se agota con estas dos perspectivas. Existe otra tendencia que construida sobre la asunción de covarianza condicional entre pares de ítems puede incluirse dentro del conjunto de modelos de respuesta al ítem no paramétricos. Esta es la base de DIMTEST y DETECT. El primero evalúa la unidimensionalidad esencial de datos binarios, a saber, la presencia de un factor dominante responsable de las respuestas observadas (Stout, 1990). El segundo es un procedimiento exploratorio que estima el número de dimensiones latentes dominantes, identifica clusters dimensionalmente homogéneos para cada dimensión y cuantifica la multidimensionalidad presente en los datos (Zhang y Stout, 1999).

Funcionamiento diferencial del ítem

La importancia de garantizar la equidad en el proceso de medición, implícita en el concepto de validez, es el origen de la multitud de trabajos destinados tanto a la elaboración y estudio de técnicas diseñadas para la detección del funcionamiento diferencial del ítem (FDI) (Camilli y Shepard, 1994; Holland y Wainer, 1993), como a la búsqueda de teorías explicativas que analicen sus causas (Hambleton, Clauser, Mazor, y Jones, 1993).

La presencia de funcionamiento diferencial en un ítem supone que la probabilidad de respuesta correcta no depende únicamente del nivel del sujeto en el espacio latente medido, sino que ésta se haya además condicionada por la pertenencia a un determinado grupo social, cultural, lingüístico, instruccional…, que genera una falta de equivalencia métrica entre puntuaciones. Su detección se apoya en procedimientos estadísticos que comparan las respuestas de sujetos que proviniendo de diferentes grupos (referencia y focal) presentan el mismo nivel en el rasgo medido. Es posible agruparlos en función del carácter observado o latente de la variable sobre la que se comparan las respuestas. La utilización de puntuaciones empíricas como criterio de equiparación de sujetos da lugar a los procedimientos conocidos como Delta, chi-cuadrado, Mantel-Haenszel, estandarización, modelos log-lineales, SIBTEST y regresión logística. Dentro del segundo grupo se incluyen los procedimientos derivados de los modelos de respuesta al ítem, y del modelo factorial. Entre todos ellos, el estadístico Mantel-Haenszel (MH) es el que mayor difusión ha alcanzado. Es un procedimiento simple para el estudio de tablas de contingencia que compara la igualdad/diferencia en la plausibilidad de la respuesta entre grupos en función del nivel de los sujetos en la variable medida.

A pesar de la estrecha relación entre los conceptos de sesgo y FDI, es importante anotar que no existe correspondencia biunívoca entre ambos. Aunque consideremos los índices de FDI definiciones operacionales del sesgo, el (in)cumplimiento de las condiciones empíricas que en cada caso exigen los procedimientos de estimación son el origen de falsas detecciones (errores tipo I) que pueden llevarnos a conclusiones erróneas. Es menester complementar todo estudio empírico de detección de FDI (Elosua, López, y Torres, 2000) con procedimientos de juicio e inferenciales que en cada caso evalúen y contextualicen los resultados antes de concluir la presencia o ausencia de sesgo. La detección estadística del funcionamiento diferencial del ítem no es un fin en sí mismo, es un instrumento útil que adquiere relevancia dentro de un marco sustantivo de estudio de la validez.

Fuentes de evidencia externas

Relaciones con otras variables

El estudio de las relaciones entre la medida obtenida por el test y variables externas, conocida como el aspecto externo de la validez por Loevinger (1957), o como amplitud nomotética por Embretson (1983), tal vez sea el tipo de evidencia más utilizado en el proceso de validación. Su defensa como fuente de validez por el enfoque funcionalista en la construcción de tests ha avalado su uso desde los primeros estándares de la APA.

Esta fuente de información se nutre de evidencias que relacionan la puntuación con algún criterio que se espera pronostique el test, con otros tests que hipotéticamente midan el mismo constructo, constructos relacionados o constructos diferentes (AERA, APA y NMCE, 1999). Los resultados de estos análisis servirían para evaluar el grado en que las relaciones hipotetizadas son consistentes con la interpretación propuesta. Este aspecto de la validez integra la evidencia convergente/discriminante, las relaciones test/criterio y los estudios de generalización de la validez, que ya en los estándares de 1985 ocuparon un apartado independiente.

Evidencia convergente y discriminante

Una de las características, y no por ello deseable, de la medición psicológica clásica es la dependencia entre la medida obtenida y el instrumento utilizado. El alcance de esta supeditación se ha estudiado habitualmente a través de la matriz multirrasgo/multimétodo (Campbell y Fiske, 1959). Su objetivo es evaluar la convergencia o divergencia esperada entre las correlaciones obtenidas en la medición de una/s variable/s por método/s diferente/s. La va-

lidez convergente (valores monorrasgo-heterométodo) se refiere al grado de relación entre distintos procedimientos que miden el mismo constructo, mientras que la validez discriminante (valores heterorasgo-monométodo) hace referencia a la evaluación de distintas variables medidas con el mismo método. Aunque en primera instancia se trate de un procedimiento heurístico se están proponiendo modelos interesantes para su estudio derivados del análisis factorial confirmatorio (Browne, 1984; Marsh y Bailey, 1991).

Relaciones test-criterio

El análisis de las relaciones test-criterio adquiere una gran relevancia en contextos de utilidad donde es fundamental la precisión con que se efectúa una predicción. Su estudio incluye la evaluación de los factores que inciden en la relación estadística entre dos o mas variables. Entre ellos las características propias del instrumento evaluado, el tamaño muestral, la restricción del rango, o la relevancia, fiabilidad y validez del criterio que se quiere pronosticar.

Los diseños utilizados para la obtención de índices de validez, propios de este aspecto, dependen del tiempo transcurrido entre la recogida de datos en el test y en el criterio, siendo habitualmente conocidos como predictivo, concurrente o retrospectivo.

En función tanto del número de variables empleadas como de su carácter sería posible la utilización de prácticamente la totalidad de técnicas de análisis multivariado, que podríamos incluir bajo el término genérico de modelo lineal generalizado. Entre ellas, regresión y correlación simple (un test / un criterio), regresión múltiple, regresión logística, análisis discriminante (varios predictores / un sólo criterio) o la correlación canónica y el análisis de regresión multivariante para el caso de varias variables predictoras y varios criterios. La aplicación de estas técnicas en los estudios de validez pueden consultarse en las obras de Martínez Arias (1995), Muñiz (1998), Paz (1996) o Santisteban, (1990) .

Generalización de la validez

La posibilidad de que los procesos de validación locales puedan extenderse a nuevas situaciones está ya reconocida en los estándares desde 1985. El objetivo es la generalización de resultados sin necesidad de nuevos estudios de validación. La base de la generalización está constituida por los estudios de meta-análisis, que en este ámbito cumplen dos objetivos complementarios. Por un lado, unificar los resultados de aplicaciones particulares de un mismo test, y por otro, estimar la variabilidad de los resultados locales obtenidos debidos a artefactos estadísticos. Hunter y Schmidt (1991) diferencian los siguientes artefactos que sería necesario neutralizar en los estudios de validación: los errores de medida, la dicotomización, la variación en el rango y la validez de constructo tanto de las variables independientes como dependientes, la varianza debida a factores extraños, el error muestral y los errores de informe o transcripción.

En definitiva se trata de estimar un promedio de validez corrigiendo los efectos de cada uno de los factores mencionados, que en nuestro entorno todavía no ha adquirido un lugar propio en la investigación psicométrica aplicada.

Consecuencias

Citada por primera vez en la revisión de 1999 tras un debate sobre su adecuación, la validez consecuencial es la fuente de evidencia más controvertida. La discusión no se ha centrado en la necesidad de evaluar las consecuencias del uso de un test, punto en el que todos los autores están de acuerdo, sino en la consideración de ésta como parte integrante de un estudio de validez. Los teóricos que más se oponen a esta perspectiva (Meherens, 1997; Popham, 1997) opinan que entremezclar ambos aspectos, pertinencia de la inferencia y consecuencias del test, enturbia excesiva e innecesariamente el significado de validez, que se ha de centrar en la justificación de la inferencia sobre una puntuación, independientemente de qué se haga con ella.

La integración del test con sus consecuencias en los estudios de validación ha sido especialmente defendida por Messick (1989). Este autor propone un marco teórico en el que integra un componente pragmático con el que enfatiza: a) la importancia de la relación entre la connotación teórica y las connotaciones prácticas atribuidas a las puntuaciones; b) la necesidad de valorar la relevancia y la utilidad de las puntuaciones en cada uno de los usos propuestos; c)la necesidad de conocer y en su caso controlar las consecuencias sociales del uso propuesto. Con ello se busca un equilibrio entre el valor instrumental del test o su finalidad y los efectos derivados de su uso, que sólo se consigue haciendo al usuario cómplice y responsable del valor terminal del test.

La postura adoptada por los estándares al respecto remarca la diferencia entre la evidencia relevante a la validez, y la evidencia que aunque relacionada con decisiones sobre las puntuaciones cae fuera de los límites de un estudio de validez. De esta suerte, la validación de un instrumento ha de considerar el análisis de la posible infrarrepresentación del constructo o de la existencia de componentes irrelevantes para el mismo; aspectos que pueden ser detectados a veces, como consecuencia del uso del test. No olvidemos que son precisamente las consecuencias sociales derivadas de un uso indiscriminado de los tests la raíz de un área de estudio psicométrico de especial relevancia social relacionada con la equidad en el proceso de medición.

Discusión

En definitiva, el proceso de validación aglutina un conjunto de estudios encaminados a proveer a las puntuaciones del test de una interpretación teórica coherente con relación a un contexto de uso bien delimitado. Es un análisis que se inicia en el momento previo a la construcción del instrumento, y que guía y acompaña su desarrollo y vigencia, asegurando interpretaciones sostenidas por un cúmulo suficiente de evidencias que garanticen equidad tanto en la administración como en la puntuación.

El concepto de validez se torna con esta definición amplio y complejo, tanto como la comprobación de teorías científicas con la que se equipara. Esta similitud aceptada y asumida por toda la comunidad psicométrica tiene una doble incidencia, aplicada y teórica, que nos gustaría resaltar. La equivalencia por un lado, convierte a los estudios de validez en áreas imprecisas. La comprobación de la validez del uso de un instrumento carece de un límite inferior objetivamente determinable, y como es lógico, es imposible fijar una cota superior. Por otro, la analogía hace referencia a un aspecto defendido desde la medición representacional, que fue olvidado desde una perspectiva operacional fuertemente arraigada en psicología, la importancia de la teoría en la medida, o la vinculación directa entre la puntuación y su significación psicológica.

El binomio puntuación-interpretación explicita además que siendo la validez uno de los pilares sobre los que se asienta un test, no es el único. La interpretación psicológica de una puntuación no puede sostenerse sin una representación formal rigurosa; aspecto del que se ocupan los modelos psicométricos (Fig. 1). Estos no son sino modelos matemáticos para la estimación de puntuaciones (V, q) que contemplan entre otros aspectos el error de medida. Sus estimaciones son la base sobre la que opera el componente de representación sustantiva aportando la significación psicológica necesaria para demarcar el valor de las inferencias.

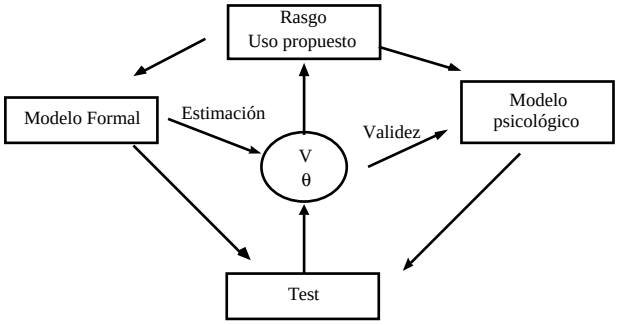

Figura 1. *Construcción de tests*

La conjunción entre ambos ejes de la medición se torna más evidente si cabe en las áreas de investigación psicométrica actual, donde todavía quedan por solucionar aspectos tanto formales como sustantivos. La incorporación al panorama educativo de la medición auténtica, los tests adaptativos informatizados, la generación automática de ítems, la utilización de Internet como medio para la creación/aplicación de cuestionarios de evaluación, o la aplicación de las últimas innovaciones multimedia al proceso de construcción de ítems, están añadiendo nuevos matices a las áreas psicométricas tradicionales.

La medición auténtica ha de solucionar problemas de representación formal relacionados por un lado con, qué y cómo puntuar, qué criterios aplicar y el modo de hacerlo… (Clauser, 2000) y por otro, con la intervención de nuevas fuentes de error de las que se hace eco la teoría de la generalizabilidad (Brennan, 2000). Además debe de responder a la posible falta de representatividad de una sola tarea y a la elevada validez heterométodo del diseño que utiliza.

La presentación de un número reducido de ítems en los tests adaptativos informatizados, independientemente de cuestiones de arranque, selección o parada intensifica los problemas referidos a aspectos de validez interna. El estudio de la relevancia o el funcionamiento diferencial (Zwick, 2000) adquieren una trascendencia mayor que en la medición tradicional, pues a medida que se reduce el número de ítems sus efectos sobre la estimación final se acentúan. Desde una perspectiva aplicada tienen que vencer la falta de validez aparente de tests que por individualizados y por tanto diferentes, son percibidos como incompletos.

Las nuevas aplicaciones multimedia, que posibilitan la construcción de ítems complejos (música, sonido, movimiento, animación…), abren la puerta al estudio y evaluación de nuevos mecanismos y acciones de respuesta (Parshall, Davey y Pashley, 2000). Es un campo de trabajo todavía virgen, en proceso de estudio, y del que se tendrán que valorar las aportaciones que suponen y acarrean a la medición tradicional.

Los problemas planteados por la teleevaluación vuelven a reflejar la interconexión entre los pilares apuntados, puntuación-representación. Aunque en los cuestionarios distribuidos por Internet los ítems utilizados corresponden en su mayoría a formatos tradicionales, este medio se ha planteado nuevos problemas relacionados fundamentalmente con la calidad de la muestra, el cuestionable anonimato de los participantes, la falta de credibilidad de muchos de ellos o la ausencia de control sobre la situación de administración que dificulta verificar la correcta comprensión de las instrucciones o las condiciones en que el participante en la investigación responde a la prueba. Son todos ellos aspectos que repercuten directamente en la calidad del dato recogido y consecuentemente en las inferencias y generalizaciones que de ellos se derivan.

A este panorama general habría que añadir la utilización de sistemas expertos para la corrección de ítems abiertos (Bennet y Bejar, 1999) o la generación automática de ítems (GAI) (Béjar, 1990) a partir de un modelo teórico propuesto. Ambos son instrumentos que se perfilan como útiles en la mejora de la calidad de la evaluación psicopedagógica, aunque todavía se están valorando la influencia que los algoritmos utilizados en la corrección automatizada ejercen sobre la instrucción, los efectos de la generación de ítems sin modelo en la construcción de tests, o los problemas de la estimación de parámetros sin muestra.

En el siglo transcurrido entre la publicación del primer test de Binet-Simon y la incorporación de los avances tecnológicos más recientes, las exigencias científicas y éticas demandadas a la dupla puntuación-significación han evolucionado en la búsqueda de una medición precisa y sustantiva. Los últimos desarrollos de los modelos formales, y la importancia otorgada a los requerimientos de validez dan fe del empeño en una medición equitativa y significativa. Las consecuencias de la irrupción de nuevas perspectivas en el panorama psicométrico con el objetivo de mejorar la calidad de la medición actual serán objeto de estudio y discusión las próximas décadas. Aunque nuestro entorno es todavía ajeno a ellas, como bien apunta Bennet (1999) para que produzcan los frutos augurados, habrán de estar siempre guiadas por una sólida fundamentación teórica. Sólo ésta podrá marcar las vías técnicas y éticas para un correcto desarrollo que será probablemente recogido en una futura edición de los estándares para el uso de los tests.

Agradecimientos

Trabajo financiado por el Ministerio de Ciencia y Tecnología dentro del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica. BSO2002-00490.

Referencias

American Psychological Association, American Educational Research Association, y National Council on Measurement in Education. (1954). Technical recommendations for psychological test and diagnostic techniques. *Psychological bulletin, 51*(2, Pt.2).

American Psychological Association, American Educational Research Association, y National Council on Measurement in Education (1966, 1974, 1985, 1999). *Standards for educational and psychological test y manuals*. Washington, DC: American Psychological Association.

Bejar, I.I. (1990) A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14*, 237-245.

Bennet, R.E. (1999) Using new technology to improve assessment. *Educational Measurement: Issues and Practice 18*(3), 5-12.

Bennet, R.E. y Béjar, I.I. (1999) Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practices, 17*, 9-17.

Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24(4),* 339-353.

Browne, M.W. (1984). The descomposition of multirait-multimethod matrices. *British journal of mathematical and statistical psychology, 37*, 1-21.

Camilli, G. y Shepard, L.A. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage.

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

Clauser, B.E. (2000). Recurrent Issues and Recent Advances in Scoring Performance Asessments. *Applied Psychological Measurement, 24(4),* 310-324.

Cronbach, L.J. (1984). *Essentials of psychological testing* (4ª ed.). New York: Harper.

Elosua, P. y López, A. (2002) Indicadores de dimensionalidad para ítems binarios. *Metodología de las Ciencias del Comportamiento 4(1),*121-137.

Elosua, P., López, A. y Torres, E. (2000). Desarrollos didácticos y funcionamiento diferencial de los ítems. Problemas inherentes a toda investigación empírica sobre sesgo. *Psicothema, 12(2)*, 198-202.

Embretson, S.E. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93(1)*, 179-197.

Embretson, S.E. (1997). Multicomponent response models. En W.J.v.Linden y R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-321). New York: Springer.

Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.

Hambleton, R.K. (1980). Test score validity and standard-setting methods. En R. A. Berk (Ed.), *Criterion-referenced measurement: the state of the art* (pp. 80-123). Baltimore: Johns Hiokins University Press.

Hambleton, R.K., Clauser, B.E., Mazor, M. y Jones, R. (1993). Advances in the detection of differentially functioning test items. *European journal of psychological assessment, 9(1)*, 1-18.

Hambleton, R.K. y Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied psychological measurement, 10*, 287-302.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.

Holland, P.W. y Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates.

Hunter, J.E. y Schmidt, F.L. (1991). Meta-analysis. En R.K. Hambleton y J.N. Zaal (Eds.), *Advances in Educational and Psychological Testing: Theory and Applicaations* (pp. 157-184). Boston: Kluwer Academic Publishers.

Jensen, A.R. (1980). *Bias in mental testing.* New York: Free Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports (Monograph Supp. 9), 3*, 635-694.

Marsh, H.W. (1988). Multirait-multimethod analysis. En J.P. Keeves (Ed.), *Educational Research, methodology and measurement. An international Handbook* . Oxford: Pergamon Press.

Martínez Arias, R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos.* Madrid: Síntesis, S.A.

McDonald, R.P. (1999). *Test theory. A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Meherens, W.A. (1997). The consequences of consequential validity. *Educational measurement: Issues and Practice, 16*, 16-19.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1.012-1.027.

Messick, S. (1989). Validity. En R.L. Linn (Ed.), *Educational Measurement* (Third Edition ed., pp. 13-104). New York: American Council on Education; Macmillan Publishing Company.

Muñiz, J. (1998). *Teoría Clásica de los tests*. (6ª ed.). Madrid: Pirámide, S.A.

Parshall, C.G., Davey, T. y Pashley, P.J. (2000) Innovative item types for computerized testing. En van der Linden y C.A.W. Glas (Eds). *Computerized adaptive testing. Theory and Practice.* (pp.129-148) Dordrecht, The Netherlands: Kluwer Academic Publishers.

Paz, M.D. (1996). Validez. En J.Muñiz (Ed.) *Psicometría* (pp.49-103). Madrid. Universitas

Popham, W. J. (1997). Consecuential validity: Right concern-wrong concept. *Educational measurement: Issues and practice, 16*, 9-13.

Prieto, G. y Delgado, A.R. (1999). Medición cognitiva de las aptitudes. En J. Olea , V. Ponsoda y G. Prieto (Eds.), *Tests informatizados. Fundamentos y aplicaciones* (pp. 207-226). Madrid: Pirámide.

Santisteban, C. (1990). *Psicometría. Teoría y práctica en la construcción de tests*. Madrid: Ediciones Norma, S.A.

Sireci, S.G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5(4)*, 299-321.

Snow, R.E. y Lohman, D.F. (1993). Cogntitive Psychology, New Test Design and new test theory: An introduction. En N. Frederiksen , R.J. Misley e I.I. Béjar (Eds.), *Test theory for a new generation of tests* (pp. 1-18). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55*, 293-326.

Zhang, J. y Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64(2)*, 213-249.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. En W. van der Linden y C.A. W. Glas (Eds.), *Computerized Adaptive Testing. Theory and Practice* (pp. 221-243). Dordrecht, The Netherlands: Kluwer Academic Publishers.