LA VALIDEZ EN LOS TESTS, ESCALAS Y CUESTIONARIOS

Juana Gómez Benito; María Dolores Hidalgo [1] [2]

Abstract

This paper seeks to highlight the need for checking the validity metrics as a characteristic fundamental to the proper use of any test, scale or questionnaire. To do so, first, sets out the historical significance of the validity of a measuring instrument from a differenciated vision of the different types of possible evidence to a unitary concept that subsumes the construct validity of any evidence that would help convergent to make the correct inferences and interpretations of the scores are obtained by applying a test and establish relations with the object of measurement constructs, also addresses the current debate surrounding the consideration or not, as part of the validity of the social consequences of the use of tests. Secondly, outlines various types of evidence that are complementary sources of validity in the sense integrator defended by Messick. Finally, it briefly addresses two hot topics in the analysis of the validity of the test, such as the detection of possible bias on the part of its items with respect to any relevant variable (eg, ethnicity, gender, culture, etc.) and the study of the metric equivalent of two different versions of the same instrument.

---

[1] Facultad de Psicología. Universidad de Barcelona. María Dolores Hidalgo Montesinos Facultad de Psicología. Universidad de Murcia Dirección postal del primer autor: Juana Gómez Benito, Departament de Metodologia de les Ciències del Comportament. Facultat de Psicologia. Universitat de Barcelona. Passeig Vall d'Hebron, 171. 08035-Barcelona (España). E-mail: jgomez@psi.ub.es
[2] Esta investigación es parte del proyecto número BSO2001-3751-C02-02 financiado por el Ministerio de Ciencia y Tecnología y la FEDER.

Resumen

El presente trabajo pretende poner en evidencia la necesidad de la comprobación de la validez como característica métrica fundamental para el correcto uso de todo test, escala o cuestionario. Para ello, en primer lugar, se expone el recorrido histórico de la significación de la validez de un instrumento de medida desde una visión disgregada de los distintos tipos de evidencias posibles a un concepto unitario que subsume en la validez de constructo cualquier evidencia convergente que ayude a realizar las inferencias e interpretaciones correctas de las puntuaciones que se obtengan al aplicar un test y establecer las relaciones con los constructos objeto de medida; se aborda también el debate actual en torno a la consideración o no, como parte de la validez, de las consecuencias sociales del uso de los tests. En segundo lugar, se delinean distintos tipos de evidencia que constituyen fuentes complementarias de validez en el sentido integrador defendido por Messick. Por último, se abordan brevemente dos temas candentes en el análisis de la validez del test, como son la detección del posible sesgo de parte de sus ítems con respecto a alguna variable relevante (como por ejemplo, etnia, género, cultura, etc.) y el estudio de la equivalencia métrica de dos versiones distintas del mismo instrumento.

Perspectiva histórica del concepto de validez

En las ciencias sociales y humanas tales como la Sociología, Psicología o la Educación se utilizan como instrumentos de medida tests, escalas, encuestas, cuestionarios y/o autoinformes, con la finalidad de obtener información acerca de opiniones, intereses, actitudes, habilidades, etc.. La comprobación de las características métricas de dichos instrumentos se constituye como la cuestión básica para decidir el uso o no de los mismos en un contexto aplicado. En este sentido es importante exigir a los cuestionarios y tests que sean fiables (precisos) y válidos. La segunda de las características métricas comentadas, la validez, resulta

una temática de máxima importancia en el proceso de construcción de un test o un cuestionario y, genéricamente hablando, requiere comprobar la utilidad de la medida realizada, es decir, el significado de las puntuaciones obtenidas. Es precisamente la validez la que permitirá realizar las inferencias e interpretaciones correctas de las puntuaciones que se obtengan al aplicar un test y establecer la relación con el constructo/variable que se trata de medir.

El concepto de validez ha ido modificándose a lo largo del tiempo, en las siguientes líneas se profundiza en ese recorrido histórico de la misma. En un principio la validez de un test psicológico, educativo o sociológico, era evaluada por una diversidad de procedimientos. El tipo de evidencia utilizado para demostrar la validez del test variaba con el propósito del test, la orientación teórica del test y el tipo de datos disponibles (Anastasi, 1986). Se hablaba de validez aparente, validez intrínseca, por definición, validez lógica, validez factorial, etc. Esta multitud de denominaciones llevó a la American Psychological Association (APA) a publicar a mediados de los años cincuenta (1954) un manual con recomendaciones técnicas para el uso y diagnóstico con tests psicológicos. En esta primera publicación de lo que más tarde constituirían los Standards se pretendía establecer algún orden en la práctica de la construcción de tests. Las consecuencias que tuvo para la validez fueron inmediatas ya que se delimitaron y denominaron los distintos tipos de validez que un test debe incluir, refiriéndose a la validez de contenido, predictiva, concurrente y de constructo. Y en general, la validez se define como el grado en el que el test mide lo que pretende medir o sirve para el propósito por el que ha sido construido. A finales de estos años cincuenta el concepto clásico-estático de validez se modifica, como bien queda recogido en las ediciones de las recomendaciones técnicas del uso de los tests (Standards) de esta época (APA, 1954) y mediados de los sesenta (APA, 1966). En los estándares de 1974, que la APA publica en colaboración con la American Educational Research Association (AERA) y el National Council on Measurement in Education (NCME), la validez predictiva y

concurrente se subsumen en criterial o de criterio. Queda así configurada la estructura tripartita de la validez (contenido, criterio y constructo), que Guion (1980) en un tono de humor definió como la visión trinitaria de la validez, y que aún hoy en día domina el campo aplicado de validación de los tests. Sin embargo, esta concepción acerca de la validez ha supuesto dos peligros para el constructor del test: i) pensar que se trata de tres tipos de validez distintos y ii) pensar que en un estudio de validez es imprescindible recoger evidencia sobre los tres aspectos de la misma.

En todo este camino recorrido por la validez, es la de constructo la que va cobrando mayor importancia. Hacia mediados de los años cincuenta, Cronbach y Meehl (1955) publican un trabajo sobre este tema, entendiendo que es uno de los aspectos más importantes, donde la comprobación de la validez de constructo implica la recogida de distintas evidencias, no todas de ellas cuantitativas, y la integración de la información recogida. La idea que se empieza a fraguar es que el proceso de validación no difiere del proceso de construcción de teorías científicas y algunas de las estrategias para investigar la validez de constructo tienen el mismo fin. Cronbach y Meehl (1955) hablan de estudiar diferencias entre grupos, correlaciones entre subtests, estudios sobre la estructura interna del test, estudios factoriales, correlaciones con criterios externos, estudios longitudinales, estudios experimentales y estudios sobre el análisis del contenido. La aportación de Campbell y Fiske (1959) introduciendo los conceptos de validez discriminante y convergente, y la matriz multimétodo-multirrasgo para evaluarlas, suponen un paso más en la importancia de este tipo de validez. Messick (1975) defiende que el concepto de validez de constructo es un concepto más general que los de validez predictiva o concurrente, que son específicos de los criterios externos utilizados así como de los grupos evaluados, y que la validez de contenido no es una propiedad de las respuestas obtenidas sino del test construido. Messick concluye que en la medida, toda la evidencia de validez debe ser de constructo. Sin embargo, los

cambios importantes en la concepción de la validez se producen a principios de los años ochenta con las aportaciones de Cronbach (1988), Guion (1980), Linn (1980) y Messick (1980) (entre otros), cuyas posiciones se recogen en el libro editado por Wainer y Braun (1988) producto de una conferencia sobre validez. Los Standards de 1985 mencionan ya que los distintos tipos de validez (criterial, de contenido, factorial, discriminante, etc.,) son distintas formas de expresión de la validez de constructo, y cualquiera de las primeras contribuye a expresar parte de la última. También recogen otra idea gestada durante estos años, a saber, lo validado no es el test mismo sino una interpretación de los datos obtenidos por un determinado procedimiento, por lo que la validez de las puntuaciones del test deben ser establecidas en cada uso que se haga del mismo. Sin embargo, el cambio drástico en la concepción de la validez se produce en los inicios de los años 80. Durante estos años numerosos trabajos enfatizan la importancia de la validez de constructo, siendo la esencia misma de todo proceso de validación, además se apunta que la validez es una, y que no podemos hablar de distintos tipos de validez, sino que todo es validez de constructo. La evolución que sufre la validez durante los años ochenta se refleja en la definición dada en la edición de 1985 de los Standards of educational and psychological testing "La validez es la consideración más importante en la evaluación de un test. El concepto se refiere a la adecuación, significado y utilidad de las inferencias específicas hechas con las puntuaciones de los tests. La validación de un test es el proceso de acumular evidencia para apoyar tales inferencias. Una variedad de evidencias pueden obtenerse de las puntuaciones producidas por un test dado, y hay muchas formas de acumular evidencia para apoyar una inferencia específica. La validez, sin embargo, es un proceso unitario. Aunque la evidencia puede ser acumulada de muchas formas, la validez se refiere siempre al grado en que esa evidencia apoya las inferencias que se hacen a partir de las puntuaciones" (APA, AERA, NCME,1985, p. 8).

La posición de Messick en cuanto a la validez contempla un aspecto más referido a

los valores sociales y las consecuencias éticas. Messick (1989) afirma que la validez no es del test o de la observación, lo que se valida son las inferencias derivadas de las puntuaciones del test o de otros indicadores, inferencias sobre el significado de las puntuaciones o la interpretación para propósitos aplicados y sobre las implicaciones para la acción, es decir, las consecuencias sociales y éticas.

El debate actual se centra en la cuestión de si la investigación de las posibles consecuencias de la administración y uso de los tests debe incluirse como una parte más del plan de validación de un instrumento, es decir, si la validación es un proceso científico o sociopolítico (Crocker, 1997). Este es precisamente el contenido del número especial de la revista Educational Measurement: Issues and Practices (vol. 16, nº 4, 1997). Las posturas están encontradas y, aunque todos asumen o destacan la importancia de las consecuencias sociales del uso de los tests, disienten en si deben ser valoradas como parte de la validez del test y uso del mismo o, por el contrario, deben ser valoradas por aquellos que desarrollan los tests pero no incluidas en la validez del mismo. Entre los defensores de la primera postura cabría citar a Linn (1997), Moss (1995), Shepard (1997) y por supuesto al propio Messick, y de la segunda, a Popham (1997) y Mehrens (1997). Para Mehrens (1997), un tema es el uso de un instrumento para llevar a cabo una medición y la precisión de la medida obtenida, y otro tema muy distinto las consecuencias que se obtengan de esa medida. Por ejemplo, un físico realiza una medición de la temperatura de un individuo y obtiene 38º C. Una cuestión es la inferencia de que ese individuo tiene fiebre y otra distinta, que hay que separar, se refiere a las consecuencias, es decir, a la decisión del tratamiento que debe seguir. Una exposición más detallada de este proceso histórico queda recogida en los trabajos de Anastasi (1986), Angoff (1988), Haney y Madaus (1991), Messick (1989), Shepard (1993, 1997).

En resumen, de las posturas actuales acerca de la validez de los tests hay que

destacar los siguientes puntos: i) lo que se valida no es el test sino las puntuaciones del test, y por lo tanto la pregunta que tratamos de responder es ¿es válido el uso o la interpretación de las puntuaciones de este test?, ii) la validez no se puede resumir en un sólo indicador o índice numérico, al igual que ocurría con la fiabilidad (coeficiente de fiabilidad, error de medida, función de información,etc.), sino que la validez de las puntuaciones de un test se asegura mediante la acumulación de evidencia teórica, estadística, empírica y conceptual del uso de las puntuaciones, iii) una puntuación puede ser válida para un uso y no para otro, iv) la validación es un proceso continuo y dinámico y v) la teoría juega un papel muy importante como guía tanto del desarrollo de un test como de su proceso de validación.

Tipos de evidencia

Aunque Messick (1980, 1989) aboga por un concepto unitario de validez, y esta concepción ha sido adoptada por la comunidad científica, como queda recogido en los últimos estándares publicados (AERA, APA, NCME, 1999), él también señaló que diferentes tipos de inferencias con los tests requieren distintos tipos de evidencia. Estos tipos de evidencia pueden obtenerse estudiando el contenido del test en función de los contenidos del dominio de referencia, examinando las relaciones entre las respuestas a las tareas, ítems y/o partes del test, estudiando las relaciones entre las puntuaciones del test y otras medidas , investigando las diferencias a través de los grupos o sobre el tiempo, y estudiando las respuestas de los sujetos a tratamientos experimentales, entre otras aproximaciones.

En general, Messick (1989, 1995) señala como aspectos a considerar en la validez:

- Contenido: relevancia y representatividad del test
- Sustantivo: razones teóricas de la consistencia observada de las respuestas
- Estructural: configuración interna del test y dimensionalidad
- Generalización: grado en que las inferencias hechas a partir del test se

pueden generalizar a otras poblaciones, situaciones o tareas. Este aspecto tiene especial importancia en la adaptación y/o traducción de escalas y tests de una cultura a otra.

- Externo: relaciones del test con otros tests y constructos. Análisis de la utilidad de la medida.
- Consecuencial: consecuencias éticas y sociales del test. Evaluación del sesgo del test.

Validez y DIF

Establecer la validez de un test implica también obtener evidencia de que el instrumento con el que se trabaja está libre de sesgo, es decir, los ítems del test funcionan del mismo modo para distintos grupos en función de variables sociodemográficas, cognitivas o de cualquier otro tipo que pueda constituir una fuente sistemática de variación ajena al constructo medido por el test. Desde un punto de vista más técnico, existirá Funcionamiento Diferencial del Item (Differential Item Functioning, DIF) con respecto a una determinada variable cuando sujetos con idéntico nivel en la característica medida por el test tengan distintas probabilidades de respuesta en el ítem dependiendo del grupo de pertenencia en la variable analizada. La detección de ítems con posible DIF es una tarea importante que debe abordarse desde los primeros momentos de elaboración de un test, y asegurar que los ítems del test están libres de DIF es un paso más en el proceso de validación del test.

La evaluación del sesgo del ítem y/o del test, así como la detección del DIF y/o del DTF (Differential Test Functioning, Funcionamiento Diferencial del Test) es una de las áreas que más investigación e interés político y social ha suscitado en la última mitad del siglo XX, existiendo en este momento una extensa variedad de técnicas de análisis para detectar DIF que se encuentran recogidas en los trabajos de Berk (1982), Fidalgo (1996), Gómez e Hidalgo (1997), Hidalgo y Gómez (1999),

Hidalgo y López (2000) Howard y Wainer (1993), Millsap y Everson (1993) y Potenza y Dorans (1995).

Validez y adaptación de escalas

La adaptación/traducción de escalas es una práctica bastante habitual por parte de sociólogos, psicológos y educadores. La Comisión Internacional de Tests (International Test Commission, ITC), que viene trabajando en esta temática en los últimos años, apunta que sólo durante el año 1992 se encontraron que algunos tests desarrollados en USA fueron traducidos y adaptados a más de 50 lenguas (Oakland, Poortinga, Schlegel y Hambleton, 2001). El proceso de traducción y adaptación de un test requiere algo más que la traducción del test de la lengua origen a la lengua destino, es necesario asegurar que las puntuaciones obtenidas con el test traducido son equivalentes a las obtenidas con el test original, para alcanzar esa equivalencia hay que considerar cuatro aspectos del proceso (Hambleton, 1994): i) el contexto cultural donde se va a realizar la adaptación, ii) aspectos técnicos del propio desarrollo y adaptación del test, iii) administración del test y iv) interpretación y documentación de las puntuaciones.

En definitiva, es necesario asegurar que el instrumento de medida presenta las mismas propiedades métricas en las dos culturas (origen y objetivo), y que por lo tanto la interpretación de las puntuaciones es la misma, es decir, existe una equivalencia métrica.

Las consecuencias para la validez son inmediatas, si el test se puede generalizar de una cultura y/o lengua a otra cultura y/o lengua, se está recogiendo evidencia acerca de la validez del mismo.

Conclusiones

En este trabajo se han abordado temas relacionados con el uso y aplicación de los

tests. Así, se han comentado algunas de las posturas más recientes acerca de cómo obtener evidencia sobre la validez de los tests y dos de los temas más actuales en el campo de la medida, el problema del funcionamiento diferencial de los tests y de los ítems, y la adaptación de tests de una cultura a otra, que se conectan directamente con el proceso de validación de un test. Estos temas son de actualidad en revistas especializadas en psicometría, pero también lo son en revistas aplicadas de sociología, psicología, educación y medicina, principalmente desde su perspectiva más práctica.

Es más, tanto los teóricos como los prácticos andan en estos momentos muy preocupados sobre el buen uso de los tests y por la calidad de los mismos, lo que conlleva un gran interés en establecer normas y directrices que regulen el uso de los tests. En este sentido las asociaciones más fuertes del ámbito psicológico, educativo y de la medida (American Psychological Association, APA, http://www.apa.org; American Educational Research Association, AERA, http://www.aera.net/; National Council on Measurement in Education, NCME, http://www.ncme.org/; y International Test Commission, ITC, http://www.intestcom.org/) se encuentran ocupadas, entre otros menesteres, en definir reglas éticas para el uso de los tests y para la construcción de los mismos con la finalidad de evitar la presencia de sesgos y factores culturales, la injusticia en la obtención de puntuaciones y, por supuesto, en la toma de decisiones. Todas estas asociaciones destacan la importancia de conocer a fondo, y estar bien formado, en los conceptos claves de la medida mediante tests y en los nuevos modelos de medida. En definitiva, la formación adecuada en medición es la clave para que los prácticos que usan tests, los usen y lo hagan de forma adecuada y ética.

Referencias

American Psychological Association. (1954). Technical recommendations for

psychological tests and diagnostic techniques. Psychological Bulletin, 51 (2, Pt.2).

American Psychological Association (1966). Standards for educational and psychological tests and manuals. Washington, DC: Author.

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

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

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

Anastasi, A. (1986). Evolving concepts for test validation. Annual Review of Psychology, 37, 1-15.

Angoff, W.H. (1988). Validity: An evolving concept. En H. Wainer y H. Braun (Eds.) Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berk, R.A. (Ed.) (1982). Handbook of methods for detecting test bias. Baltimore: John Hopkins University Press.

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

Crocker, L. (1997). Editorial: The great validity debate. Educational Measurement: Isssues and Practice, 16, 4.

Cronbach, L.J. (1988). Five perspectives on validation argument. En H. Wainer y H. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Fidalgo, A.M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord.), Psicometría. Madrid: Universitas.

Gómez, J. e Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. Anuario de Psicología, 74, 3-32.

Guion, R.M. (1980). On trinitarian doctrines of validity. Professional Psychology, 11, 385-389.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-244.

Haney, W. y Madaus, S. (1991). The evolution of ethical and technical standards for testing. En R.K. Hambleton y J.N. Zaal (Eds.) Advances in Educational and Psychological Testing: Theory and Applications. Boston: Kluwer Academic Publishers.

Hidalgo, M.D. y Gómez, J. (1999). Técnicas de detección del funcionamiento diferencial en ítems politómicos. Metodología de las Ciencias del Comportamiento, 1, 39-60.

Hidalgo, M.D. y López, J.A. (2000). Funcionamiento diferencial de los ítems:

Presente y perspectivas de futuro. Metodología de las Ciencias del Comportamiento, 2, 167-182.

Holland, P.W. y Wainer, H. (Eds.) (1993). Differential Item Functioning. Hillsdale, NJ: LEA.

Linn, R.L. (1980). Issues of validity for criterion-referenced measures. Applied Psychological Measurement, 4, 547-561.

Linn, R.L. (1997). Evaluating the validity of assessments: The consequences of use. Educational Measurement: Issues and Practice, 16, 14-16.

Mehrens, W.A. (1997). The consequences of consequential validity. Educational Measurement: Issues and Practice, 16, 16-18.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1989). Validity. En R.L. Linn (Ed.), Educational Measurement (3th. Ed.). New York: American Council on Education and Macmillan publishing company.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational measurement: Issues and Practice, 14, 5-8.

Millsap, R.E. y Everson, H.T. (1993). Methodology Review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement ,17, 297-334.

Moss, P.A. (1995). Themes and variations in validity theory. Educational

Measurement: Issues and Practice, 14, 5-13.

Oakland, T., Poortinga, Y.H., Schlegel, J. y Hambleton, R.K. (2001). International Test Commission: Its History, Current Status, and Future Directions. International Journal of Testing, 1, 3-32.

Popham, W.J. (1997). Consequential validity: Right concern-wrong concept. Educational Measurement: Issues and Practice, 16, 9-13.

Potenza, M.T. y Dorans, N.J. (1995). DIF Assessment for polytomously scored items: A framework for classification and evaluation. Applied Psychological Measurement, 19, 23-37.

Shepard, L.A. (1993). Evaluating test validity. Review of Research in Education, 19, 405-450.

Shepard, L.A. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16, 5-8, 13, 24.

Traub, R.E. (1997). Classical test theory in historical perspective. Educational Measurement: Issues and Practice, 16 (4), 8-14.

Wainer, H. y Braun, H. (Eds.) (1988). Test validity. Hillsdale, NJ: LEA.