ACTA COMPORTAMENTALIA
Vol.13 , Núm.1, junio 2005, pp. 9-20

# La Validez desde una óptica psicométrica

*(Validity from a psychometric perspective)*

**José Muñiz***

Universidad de Oviedo

En el presente trabajo se ofrece una panorámica general del estado actual de la validez desde un punto de vista psicométrico. Debido a su naturaleza, la medición de las variables psicológicas conlleva una problemática especial (Muñiz, 1998), cualquiera que sea el enfoque que se utilice para evaluarlas, constituyendo la validez el concepto central de la medición. Aquí no se emprende un análisis de las implicaciones del enfoque psicométrico de la validez para las teorías psicológicas de carácter conductual, véase al respecto por ejemplo Silva (1989). Se asume de entrada que todo lo dicho sobre el proceso de validación desde el punto de vista psicométrico es en gran medida, sino totalmente, aplicable a cualquier otro enfoque psicológico que pretenda explicar y predecir la conducta humana. Un tratamiento más amplio y comprensivo que el del presente trabajo puede consultarse en Muñiz (2004).

## NOTA HISTÓRICA Y CONCEPTUAL

Empezando por lo obvio, hay que recordar que las respuestas a los tests son conductas, pudiendo definirse un test como una muestra de conducta de una persona recogida de forma objetiva y estandarizada en un momento determinado. Los psicólogos y otros profesionales recogen esas muestras de conducta porque a partir de ellas pueden hacer inferencias fundadas acerca de la conducta y funcionamiento cognoscitivo de las personas evaluadas. La primera condición para que un test sirva de base para llevar a cabo inferencias de interés es que la muestra de conducta recogida sea precisa, es decir, que los errores cometidos en la medición, en la cata, sean aceptables, a sabiendas de que ninguna medición científica está totalmente exenta de error. La tecnología psicométrica

* Dirigir toda correspondencia al autor a: Departamento de Psicología. Universidad de Oviedo. Plaza Feijoo s/n. 33003 Oviedo (España)
Correo electrónico: jmuniz@uniovi.es

desarrollada desde hace ya un siglo (Spearman, 1904a y b) para evaluar el grado de precisión de las mediciones realizadas con los tests se denomina *fiabilidad* y constituye uno de los capítulos mejor desarrollados por la psicometría. La tecnología psicométrica encargada de mostrar que las inferencias hechas acerca del funcionamiento de las personas a partir de tests fiables son correctas es lo que denominamos genéricamente *validez*. ¿De qué modo se comprueba que las inferencias hechas a partir de un test son correctas? O más sencillamente, ¿cómo se procede para llevar a cabo el proceso de validación de las inferencias hechas a partir de un test? De ello trataremos aquí, pero antes de nada hay que responder a la pregunta *¿Qué se valida?* Ha de quedar claro desde el principio que aunque se hable con frecuencia de *validar un test,* en sentido estricto no es el test lo que se valida, sino las inferencias que se hacen a partir de él sobre determinados aspectos de la conducta de las personas. Por tanto, el resultado final de un proceso de validación no es llegar a decir de forma simplista que tal o cual test es válido; las que son o no válidas son las inferencias hechas a partir del test, no el test en sí mismo. Esto es natural, pues a partir de un test pueden hacerse inferencias de muy diverso tipo, siendo unas serán válidas y otras no; el proceso de validación consistirá precisamente en aportar datos y argumentos que permitan saber cuáles de las inferencias están fundadas, cuáles son válidas. ¿Cómo se aportan esos datos validantes? Es decir, ¿cómo se allega la evidencia empírica y teórica necesaria para poder afirmar que determinadas inferencias realizadas son válidas? La respuesta a esos interrogantes es lo que constituye el meollo de la validez psicométrica. La respuesta, como no podía ser de otro modo, ha ido evolucionando a lo largo de la historia de la psicometría. Esta evolución queda muy bien reflejada en la literatura especializada, en especial en las sucesivas ediciones del manual sobre medición educativa editado por Lindquist (1951), Thorndike (1971) y Linn (1989), y, sobre todo, en las sucesivas ediciones de los estándares publicados por la APA (1954, 1966, 1974, 1985, 1999), que representan el consenso psicométrico oficial de cada época. El concepto de validez, y por ende las prácticas de validación, ha ido evolucionando desde unos inicios marcadamente empíricos y operacionales a la situación actual en la que se entiende la validez de una forma más amplia y comprensiva. Así, cuando Gulliksen (1950) sintetiza en su excelente manual lo esencial de la teoría clásica de los tests de entonces, el problema de la validez se reduce a la correlación entre el test y el criterio a predecir. De modo que la tecnología psicométrica de la validez se centraba en el estudio de las correlaciones entre el test y los criterios a predecir, y las variables que modulaban esta relación, tales como la variabilidad de la muestra utilizada, la longitud del test, la fiabilidad del test y del criterio, o determinadas covariantes. Nada que objetar, y esta tecnología clásica sigue vigente en la actualidad; lo que ocurre es que además de los datos relativos a la correlación test-criterio el concepto de validez se ha ido ampliando paulatinamente. El trabajo

pionero de Cronbach y Meehl (1955) sobre la *validez de constructo* alerta a teóricos, constructores y usuarios acerca de la importancia de ocuparse de la rigurosidad y entidad del constructo medido, además, obviamente, de trabajar con las correlaciones test-criterio. A partir de entonces, durante muchos años las vías esenciales para recoger datos en el proceso de validación de los tests fueron el análisis de los contenidos de la prueba, las correlaciones test-criterio y la entidad de los constructos, lo que dio lugar a que se hablase de la santísima trinidad de la validez: validez de contenido, validez de criterio y validez de constructo. Los estándares de la AERA, APA y NCME de 1985 dejan bien claro que si bien esas tres vías de recogida de datos son legítimas, la validez es sólo una y no hay razón alguna para que se obtengan datos por cualquier otro camino complementario. Ese es el planteamiento dominante sobre validez en los ochenta, que en el fondo no es otra cosa que subsumir los planteamientos sobre validez en el marco más general de la comprobación de hipótesis científicas. Validar las inferencias de los tests es un caso particular de la validación de modelos e hipótesis científicas. Esto no resuelve ningún problema metodológico específico en el proceso de validación; más bien hace ver a los validadores de tests que sus problemas no son muy distintos a los que tienen el resto de los científicos.

Desde que se publicaran los estándares de la AERA, APA y NCME (1985) la psicometría ha conocido grandes avances en todas las ramas y la validez no es una excepción, si bien las novedades en este campo no han sido tan espectaculares como en otros. Se mantiene la filosofía general de la validez como un planteamiento unitario, con Messick (Messick, 1980, 1988, 1989) como apóstol principal, aunque se utilicen distintas aproximaciones para obtener datos relevantes para la validación de las inferencias. Como ya se ha señalado, validar un test puede considerarse un caso particular de la comprobación de hipótesis científicas, pero no existe un método científico claro y universal (Weinberg, 2003) que aplicado de forma algorítmica dé solución a todos los problemas, lo cual tampoco quiere decir que todo vale. Este es un planteamiento correcto y teóricamente justificado, pero como señala Brennan (1998, 2001), si bien la noción de una validez unitaria es muy sugerente teóricamente, hasta la fecha no ha mostrado una gran utilidad práctica cara a los procesos reales de validación. Los constructores y usuarios de los tests reclaman reglas más específicas que les permitan allegar datos que les ayuden a validar sus inferencias. Las tres vías clásicas para la recogida de datos, a saber, la validez de contenido, de criterio y de constructo siguen siendo feraces, por supuesto, pero algunas otras se han ido añadiendo en este proceso de construcción de la validez. Así en los estándares de 1999 (AERA, APA y NCME, 1999) se señalan además de estas tres, los procesos de respuesta implicados, la problemática de la generalización de la validez, o las consecuencias del uso de los tests. Tal vez convenga dedicar unas palabras a este último aspecto de las consecuencias, el cual ha generado cierto debate entre la comunidad psicométrica.

## CONSECUENCIAS DEL USO DE LOS TESTS

El debate sobre lo que ha dado en llamarse validez consecuencial se aviva a raíz del influyente trabajo de Messick (1989), donde propone ampliar el marco conceptual de la validez para dar cabida en él a las consecuencias del uso de los tests. Su propuesta cala en la comunidad psicométrica dominante, hasta el punto de ser incluida en los últimos estándares de 1999. Bien es verdad que no hay unanimidad al respecto, siendo recomendables los trabajos de Shepard (1997) y Linn (1997) a favor, y los de Popham (1997) y Mehrens (1997) en contra. La literatura generada es abundante, véase por ejemplo el monográfico de la revista *Educational Measurement: Issues and Practice* (Green, 1998; Lane, Parke y Stone, 1998; Linn, 1998; Moss, 1998; Reckase, 1998; Taleporos, 1998). El meollo del debate se centra fundamentalmente en si es apropiado o no incluir las consecuencias sociales del uso de los tests en el marco de la validez. De lo que nadie duda es de la importancia de estas y de la necesidad de ocuparse de ellas por parte de los distintos agentes implicados en la utilización de los tests, tales como autores, constructores, distribuidores, usuarios, personas evaluadas e instituciones contratantes (Haertel, 2002; Kane, 2002; Lane y Stone, 2002; Ryan, 2002). Al incluir las consecuencias sociales en el marco de la validez se corre el riesgo de introducir por la puerta de atrás los planteamientos éticos y políticos en el estudio de la validez, que debería reservarse para los argumentos científicos. Autores como Maguire, Hattie y Brian (1994) consideran que esta insistencia en incluir las consecuencias dentro del marco de la validez viene motivada en gran parte por las continuas refriegas legales que rodean a los tests en Estados Unidos. Consideran que si bien esta postura pudiera reducir las batallas legales, también puede distraer a los constructores de su misión central que no es otra que aportar datos de cómo el test representa al constructo medido.

Nótese que esta cuestión de la validez consecuencial no se identifica estrictamente con el uso inadecuado de los tests, éste sencillamente ha de evitarse, para lo cual las organizaciones nacionales e internacionales llevan a cabo muy diversas iniciativas; véase, por ejemplo, Bartram (1998), Fremer (1996), Evers (1996), Muñiz (1997, 1998), Muñiz y Fernández-Hermida (2000), Muñiz, Prieto, Almeida y Bartram (1999), Simner (1996). Una alternativa razonable sería incluir en esta tradición del uso adecuado de los tests todo lo relativo a las consecuencias, pero hay quien considera que esto sería rebajar la importancia atribuida a ellas, ya que incluidas en el capítulo de la validez tienen garantizada una mayor relevancia en el debate psicométrico.

## OTEANDO EL FUTURO

Además de los planteamientos ya señalados, centrales en el proceso de validación, ¿qué hay realmente nuevo y emergente en relación con la validez ahora que comienza

el siglo XXI? Los retos para la validez y para la medición psicológica y educativa en general provienen de los espectaculares avances tecnológicos experimentados por los instrumentos de medida. Por ejemplo, ¿cómo validamos los nuevos tests adaptativos informatizados? ¿Cómo va a influir en la validez la nueva era de ítems multimedia? ¿Será posible validar de forma eficiente los nuevos sistemas de evaluación auténtica? ¿Cómo se controla la validez de la tele-evaluación? ¿Qué grado de validez cabe esperar de los sistemas de corrección automatizada? ¿Lograrán los sofisticados modelos de TRI potenciar la validez de forma significativa? He ahí algunos interrogantes tomados a vuela pluma, sin ningún ánimo de exhaustividad. Esas son algunas de las líneas por las que se adivinan los verdaderos retos de la validez que viene. Se comentan brevemente estos interrogantes, aunque cada uno de ellos por separado tendría entidad suficiente para dedicarle un trabajo.

## TESTS ADAPTATIVOS INFORMATIZADOS

Los tests adaptativos informatizados (TAIs) constituyen el avance tecnológico más destacado de la psicometría de los últimos años (Olea, Ponsoda y Prieto, 1999; Van der Linden y Glass, 2000; Wainer, 1990). Aportan ventajas notables para medición de las variables, tanto desde el punto de vista de la precisión, como de la economía temporal, seguridad, control del proceso evaluativo, inmediatez de los resultados, o motivación de las personas evaluadas; de ahí que sean considerados como el avance tecnológico por excelencia de la psicometría contemporánea, la aplicación *matadora.* Pero curiosamente no son demasiados los esfuerzos dedicados a analizar su validez (Muñiz y Hambleton, 1999), dándose con frecuencia por supuesta, como si la excelente precisión que alcanzan con pocos ítems fuese garantía de su validez. Nada más alejado de la realidad; como se ha señalado en las líneas precedentes, la validez ha de demostrarse explícitamente, tanto teórica como empíricamente. Los TAIs plantean muchos retos de validación desde diferentes ángulos. El planteamiento clásico de validar a partir de las puntuaciones en el mismo test ha de modificarse, pues aquí cada sujeto recibe un test distinto, si bien es verdad que las puntuaciones se expresan en la misma escala theta. El número de sujetos que responden a determinados ítems es pequeño con las consecuentes dificultades de análisis estadísticos que ello conlleva. Los modelos psicométricos utilizados habitualmente con los TAIs son unidimensionales, mientras que numerosos constructos de interés son multidimensionales. El uso del ordenador puede estar introduciendo fuentes de error no deseadas debido a la distinta familiaridad de los examinados con las máquinas (Taylor, Kirsch, Eignor y Jamieson, 1999), o con cierta plataforma (Macintosh frente a Windows), o debido a la interacción hombre-máquina. Por otro lado, los efectos de la diferente velocidad

para trabajar se acentúa con los TAIs, incrementando la ansiedad en algunos casos, ya que a diferencia de los tests de papel y lápiz aquí no se pueden repasar las respuestas y corregir. Muchos de estos aspectos están siendo abordados por los investigadores, pudiéndose consultar una buena revisión en Huff y Sireci (2001). Dada la preocupación actual por las consecuencias del uso de los tests, no está de más señalar que algunos de los programas de evaluación actual basados en los TAIs resultan mucho más caros que los tradicionales, por lo que una consecuencia indeseada es el perjuicio que esto puede causar a los desfavorecidos económicamente. En definitiva, los tests adaptativos informatizados abren todo un conjunto de retos cara a su validación rigurosa.

## ITEMS MULTIMEDIA

Otro de los puntos calientes mencionados es la validación de los ítems multimedia. En general la tecnología para la construcción de ítems ha conocido en los últimos años un desarrollo inusitado (Haladyna, 1999; Haladyna, Downing y Rodríguez, 2002; Osterlind, 1998; Moreno, Martínez y Muñiz, 2004; Prieto y Delgado, 1996). Varias son las razones para esta eclosión, resumidas por Muñiz y García-Mendoza (2002) en cinco fundamentales: a) el interés centrado en los ítems de los modelos de TRI, b) el desarrollo de los TAIs, verdaderos devoradores de ítems, c) el desarrollo de la tecnología del funcionamiento diferencial de los ítems, d) la convergencia entre la psicología cognitiva y la psicometría (Frederiksen, Mislevy y Bejar, 1993; Prieto y Delgado, 1996, 1999), e) la aparición del movimiento sobre la "evaluación auténtica" que ha servido de estímulo para aguzar el ingenio de los constructores de tests convencionales. Dentro de este contexto tan efervescente emerge un nuevo tipo de ítem que aprovecha todas las ventajas de la informática, incluyendo sonido, animación, posibilidades de interacción, realidad virtual, conexión a la red, en suma, todas las posibilidades que hoy nos ofrece el mundo de los multimedia. Aparte de las complejidades técnicas, la cuestión de fondo es si este tipo de ítems tan atractivos y prometedores van aportar un incremento de la validez, o van a quedarse en un mero cambio de soporte de los tests. A su favor tienen el gran realismo que permiten, pudiendo simular situaciones muy cercanas a la realidad cotidiana de los sujetos, lejos de la asepsia de las modestas matrices en blanco y negro. La duda, aún por resolver, es si este realismo no será precisamente su talón de Aquiles, al perder capacidad de generalización, precisamente por estar tan pegados a una situación concreta. No sabemos como evolucionarán estos tests, la tarea de validarlos está pendiente. Para buenas presentaciones de los nuevos formatos de ítems véase por ejemplo, Parshall y Balizet (2001), Parshall, Davey y Pashley (2000), o Zenisky y Sireci (2003).

## EVALUACIÓN AUTÉNTICA

Un tercer interrogante hacía alusión a la llamada evaluación auténtica (Bravo y Fernández del Valle, 2000; Hakel, 1998; Powell, 1990), movimiento surgido en los ámbitos educativos y que se refiere con el término un tanto pretencioso de *auténtica* a un tipo de evaluación educativa supuestamente más abierta y flexible, menos estandarizada, que los tests convencionales. Se trataría de acercar la evaluación a las situaciones reales, utilizando por ejemplo, los *portafolios*, que incluirían la evaluación de todo tipo de actividades realizadas en un periodo de tiempo por el alumno, tales como redacciones, libros leídos, videos, dibujos, trabajos, etc. Nada que objetar, todo muy realista, bienintencionado y pegado a la realidad del alumno, pero hay que demostrar empíricamente su validez real. ¿Es mejor predictor este tipo de pruebas que el más convencional objetivo y estandarizado? ¿Cómo se validan y a qué costo estas estrategias? ¿Cuál es la fiabilidad inter-jueces de este tipo de pruebas? Bravo y Fernández del Valle (2000) señalan varias limitaciones para la implantación generalizada de estos nuevos modelos, tales como su alto costo en tiempo y dinero, la dificultad de realizar evaluaciones paralelas, falta de acuerdo sobre los constructos a medir, fiabilidad inter-jueces, naturaleza compleja de muchos de los ejercicios, sesgos introducidos por el contexto, o pobre generalización de los resultados. Según Bachman (2002) cuatro son las cuestiones cruciales a responder a la hora de validar este tipo de evaluación: a) qué constructo se mide y qué inferencias concretas se pueden hacer sobre las aptitudes de las personas evaluadas, b) cuál es el ámbito de generalización de los resultados, c) cuál es el grado de *autenticidad* real de las respuestas, y d) cuál es el grado de interactividad de las respuestas con la tarea de evaluación. En fin, una propuesta interesante que abre seguramente más interrogantes que los problemas que viene a resolver. De nuevo nos tememos que pueda ser más caro el remedio que la enfermedad, el futuro lo irá diciendo.

## TELE-EVALUACIÓN

Otro de los interrogantes citados aludía a la tele-evaluación que empieza a surgir con fuerza. El desarrollo de los medios de comunicación en general y de Internet en particular, facilita el que se puedan llevar a cabo evaluaciones a distancia, tanto tradicionales como basadas en los tests adaptativos informatizados. Este tipo de evaluación plantea serios problemas que van de lo ético a la autentificación (asegurarse que la persona evaluada es la que responde), pasando naturalmente por el problema de la validación de las pruebas en esas circunstancias. No se sabe con certeza como afectará esta circunstancia de evaluación a distancia en los datos relativos a la evaluación.

## CORRECCIÓN AUTOMATIZADA

Otro avance técnico reciente que dará trabajo en el ámbito de la validez es la cada vez más frecuente corrección automática de ensayos y respuestas abiertas mediante programas de ordenador (Bejar y Bennet, 1999; Bennet y Bejar, 1999; Clauser, 2000; Clauser, Harik y Clyman, 2000; Martinez y Bennet, 1992). Los resultados son prometedores, logrando que algunos de los programas compitan en eficacia con los humanos, pero se abren numerosos interrogantes acerca generalización de unos campos de contenido a otros, o su capacidad para recoger contenidos atípicos o innovadores. En definitiva, que lo no programado no va a ser valorado.

## TEORÍA DE RESPUESTA A LOS ÍTEMS

Conviene recodar que los grandes avances de la psicometría contemporánea han venido de la mano de los modelos de teoría de respuesta a los ítems, centrándose fundamentalmente, aunque no sólo, en la mejora de la precisión de las mediciones. Ello es esencial, pues difícilmente se puede hablar de validez sin fiabilidad, pero en este contexto de la TRI no se ha hecho el esfuerzo paralelo en validez al realizado con la fiabilidad. Con mucha frecuencia los investigadores y usuarios parecen dar por terminada su labor cuando dan con un modelo de medición que se ajusta a los datos, y eso solo es el principio cuando de validar un modelo se trata.

Señalar finalmente, que no faltan quienes, como Borsboom y Mellenbergh (2004), proponen dar un giro radical al enfoque psicométrico de la validez, empezando de cero. Su provocadora propuesta merece ser tenida en cuenta en el debate y análisis sobre la validez, pero se sale de nuestro propósito y espacio en el presente trabajo. Algo parecido ocurre con la validación en los ámbitos educativos, muy bien planteada en el reciente trabajo de Lane (2004).

Esperamos que estas pinceladas sobre validez permitan a los lectores hacerse una idea aproximada acerca de las tendencias y problemática actual sobre la validez desde el punto de vista de la psicometría.

## REFERENCIAS

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

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

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

American Psychological Association. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*, 5-18.

Bartram, D. (1998). The need for international guidelines on standards for test use: A review of European and international initiatives. *European Psychologist, 2*, 155-163.

Bejar, I. y Bennet, R. (1999). La puntuación de las respuestas como un parámetro del diseño de exámenes: implicaciones en la validez. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: fundamentos y aplicaciones*. Madrid: Pirámide. (Pp. 53-59).

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

Borsboom, D. y Mellenbergh, G.J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Bravo, A. y Fernández del Valle, J. (2000). La evaluación convencional frente a los nuevos modelos de evaluación auténtica. *Psicothema, 12* (Supl.), 95-99.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*, 5-9.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20(4)*, 6-18.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessment. *Applied Psychological Measurement, 24*, 310-324.

Clauser, B. E., Harik, P. y Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37*, 245-261.

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

Evers, A. (1996). Regulations concerning test qualifications and test use in The Netherlands. *European Journal of Psychological Assessment, 12*, 153-159.

Frederiksen, N., Mislevy, R. y Bejar, I. (Eds.) (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: LEA.

Fremer, J. (1996). Promoting high standards for test use: Developments in the United States. *European Journal of Psychological Assessment, 12*, 160168.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17, 16-19.*

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

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21*, 16-22.

Hakel, M. D. (Ed.) (1998). *Beyond multiple choice: evaluating alternatives to traditional testing for selection*. Mahwah, NJ: LEA.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. Hillsdale, NJ: LEA.

Haladyna, T. M., Downing, S. M. y Rodríguez, M. C. (2002). A review of múltiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Huff, K. L. y Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*, 16-25.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*, 31-41.

Lane, S. (2004). Validity of high-stakes assessment: are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, 23(3), 6-14.

Lane, S., Parke, C. S. y Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17, 24-28.*

Lane, S. y Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*, 23-30.

Lindquist, E. F. (Ed.) (1951). *Educational Measurement.*(2nd edition). Washington, DC: American Council on Education.

Linn, R. L. (Ed.) (1989). *Educational Measurement.*(3rd edition). Washington, DC: American Council on Education.

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

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17,* 28-30.

Maguire, T. Hattie, J. y Brian, H. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research, 40*, 109-126.

Martínez, M. E. y Bennet, R. E. (1992). A review of automatically scorable constructed-response item types for large scale assessment. *Applied Measurement in Education, 5*, 151-169.

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

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

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. En H. Wainer y H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: LEA. (Pp.33-45).

Messick, S. (1989). Validity. En R. Linn (Ed.), *Educational Measurement*. Washington, DC: American Council on Education. (Pp. 13-103).

Moreno, R., Martínez, R. y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema, 16,* 490-497

Moss, P. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17, 6-12.*

Muñiz, J. (1997). Aspectos éticos y deontológico de la evaluación psicológica. En A. Cordero (Ed.), *Evaluación psicológica en el año 2000*. Madrid: TEA Ediciones. (Pp. 307-345).

Muñiz, J. (1998). La medición de lo psicológico. *Psicothema, 10*, 1-21.

Muñiz, J. (2004). La validación de los tests. *Metodología de las Ciencias del Comportamiento*, 5(2), 121-141.

Muñiz, J. y Fernández-Hermida, J. R. (2000). La utilización de los tests en España. *Papeles del Psicólogo, 76*, 41-49.

Muñiz, J. y García-Mendoza, A. (2002). La construcción de ítems de elección múltiple. *Metodología de las Ciencias del Comportamiento*, Monográfico, 416-422.

Muñiz, J. y Hambleton, R. K. (1999). Evaluación psicométrica de los tests informatizados. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: fundamentos y aplicaciones*. Madrid: Pirámide. (Pp. 23-52).

Muñiz, J., Prieto, G., Almeida, L. y Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment, 15*, 151-157.

Olea, J., Ponsoda, V.y Prieto, G. (Eds.) (1999). *Tests informatizados: fundamentos y aplicaciones*. Madrid: Pirámide.

Osterlind, S. J. (1998). *Constructing test items: multiple choice, constructed-response, performance, and other formats*. Boston: Kluwer Academic Publishers.

Parshall, C. G. y Balizet, S. (2001). Audio computer-based tests: a initial framework for the use of sound in computerized tests. *Educational Measurement: Issues and Practice, 2*, 5-15.

Parshall, C. G., Davey, T. y Pashley, P. (2000). Innovative item types for computerized testing. En W. J. van der Linden y C. Glass (Eds.), *Computer-adaptive testing: Theory and practice*. Boston: Kluwer Academic Publishers.

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

Powell, M. (1990). *Performance assessment: panacea or pandora's box*? Rockville, MD: Montgomery Cunty Public Schools.

Prieto, G. y Delgado, A. R. (1996). Construcción de ítems. En J. Muñiz (Ed.), *Psicometría*. Madrid: Universitas. (Pp. 105-138).

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*. Madrid: Pirámide. (Pp. 207-226).

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17, 13-16.*

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. . *Educational Measurement: Issues and Practice, 21*, 7-15.

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

Silva, F. (1989). *Evaluación conductual y criterios psicométricos*. Madrid: Pirámide.

Simner, M. L. (1996). Recommendations by the Canadian Psychological Association for improving the North American safeguards that help protect the public aginst test misuse. *European Journal of Psychological Assessment, 12*, 72-82.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1904b). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Taleporos, E. (1998). Consequential validity: A pratitioner's perspective. *Educational Measurement: Issues and Practice, 17, 20-23.*

Taylor, C., Kirsch, I., Eignor, D. y Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*, 219-274.

Thorndike, R. L. (Ed.) (1971). *Educational Measurement* (2nd edition). Washington, DC: American Council on Education.

Van der Linden, W. J. y Glass, C. (Eds.) (2000). *Computer-adaptive testing: theory and practice*. Boston: Kluwer Academic Publishers.

Wainer, H. (Ed,) (1990). *Computer adaptive testing: a primer*. Hillsdale, NJ: LEA.

Weinberg, S. (2003). *Plantar cara. La ciencia y sus adversaries culturales*. Barcelona: Paidós.

Zenisky, A. L. y Sireci, S. G. (2003). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337-362.

## RESUMEN

El estudio de la validez constituye el eje central de los análisis psicométricos de los instrumentos de medida. En esta comunicación se traza una breve nota histórica de los distintos modos de concebir la validez a lo largo de los tiempos, se comentan las líneas actuales, y se tratan de vislumbrar posibles vías futuras, teniendo en cuenta el impacto que las nuevas tecnologías informáticas están ejerciendo sobre los propios instrumentos de medida en Psicología y Educación. Cuestiones como los nuevos formatos multimedia de los ítems, la evaluación a distancia, el uso intercultural de las pruebas, las consecuencias de su uso, o los tests adaptativos informatizados, reclaman nuevas formas de evaluar y conceptualizar la validez. También se analizan críticamente algunos planteamientos recientes sobre el concepto de validez.

Palabras clave: Validez, psicometría, medida, nota histórica

## ABSTRACT

The study of validity constitutes a central axis of psychometric analyses of measurement instruments. This paper presents a historical sketch of different modes of conceiving validity, with commentary on current views, and it attempts to predict future lines of research by considering the impact of new computerized technologies on measurement instruments in psychology and education. Factors such as the new multimedia format of items, distance assessment, the intercultural use of tests, the consequences of the latter, or the development of computerized adaptive tests demand new ways of conceiving and evaluating validity. Some recent thoughts about the concept of validity are also critically analysed.

Key words: Validity, Psychometric, measurement, historical review.