Psicothema 2006. Vol. 18, nº 2, pp. 307-312
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

La evaluación de las consecuencias del uso de los tests en la teoría de la validez

José Luis Padilla, Juana Gómez*, María Dolores Hidalgo** y José Muñiz***
Universidad de Granada, * Universidad de Barcelona, ** Universidad de Murcia y *** Universidad de Oviedo

Pocas dudas plantea la importancia de la validez durante la elaboración y evaluación de los tests. Sin embargo, la incorporación de las consecuencias del uso de los tests como una fuente más de evidencia en la última edición de los estándares de la AERA, APA y NCME ha generado un intenso debate. Tras revisar los antecedentes históricos y las posturas más significativas de defensores y críticos, se exponen los argumentos para responder a la pregunta: ¿Cuándo el análisis de las consecuencias del uso de los tests debe formar parte de la validación? La respuesta de los estándares de la AERA, APA y NCME tiene claras similitudes con la planteada para la consideración del sesgo en los tests como un problema de validez. Por último, se señala la tensión que para la validación de las consecuencias puede acarrear la evolución de las nociones de justicia en el uso de los tests.

Evaluation of consequences of test use invalidity theory. There is little doubt about the importance of validity during the compilation and evaluation of tests. Nevertheless, intense debate has arisen with regard to incorporating the consequences of test use as a further source of evidence in the most recent edition of the AERA, APA, NCME *Standards*. After reviewing the historical antecedents of the issue and the main lines of approach of both defenders and critics, this paper sets out the arguments which may be used in answer to the question: 'At what point should the analysis of the consequences of test use become part of validation?' The response of the AERA, APA, NCME *Standards* has clear similarities with the perspective adopted towards the consideration of bias in tests as a problem of validity. Finally, the paper describes how consequence validation may be subject to tension arising from the evolution of notions of justice in test use.

La última edición disponible de los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1999) ha reafirmado la importancia de la validez al valorarla como «… la consideración más importante en la elaboración y evaluación de los tests» (p. 9). Hambleton y Pitoniak (2002) explican el aumento en la «visibilidad» de la validez por los retos procedentes de los últimos desarrollos en medición y evaluación, junto al uso creciente de los tests en la toma de decisiones críticas para las personas e instituciones: contratación, selección, diagnóstico, graduación, etc.

La evolución histórica y los argumentos principales del consenso actual sobre los contenidos de la teoría de la validez pueden consultarse en numerosas revisiones recientes. Elosua (2003) propone tres etapas para ordenar la evolución histórica de la teoría de la validez utilizando el criterio habitual de las sucesivas ediciones de los estándares de la AERA, APA y NCME. Kane (2001) resalta los cuatro aspectos más notables del consenso actual: a) la validez implica una evaluación de la plausibilidad de la interpretación y del uso propuesto de las mediciones; b) los juicios de validez reflejan la suficiencia y conveniencia de la interpretación propuesta y de las evidencias que la apoyan; c) la evaluación debe incluir la evaluación de las consecuencias del uso de los tests; y d) la validez es una evaluación unificada de la interpretación propuesta. Muñiz (2003), tras llamar la atención sobre las limitaciones prácticas de la noción de validez unitaria, hace una revisión exhaustiva de las vías más habituales para recoger evidencias con las que abordar el proceso de validación de las inferencias hechas a partir de un test.

Sin embargo, los retos planteados por la noción de validez unitaria han contribuido a la separación entre la teoría y la práctica de la que han alertado numerosos autores. Shepard (1993) explicaba esta separación porque la evolución de la teoría había modificado el significado de términos antiguos sin proponer nombres nuevos, y porque los requisitos de las nociones más recientes —la validez de constructo como un proceso sin fin— se perciben como demasiado exigentes. Hogan y Agnello (2004) aportan una prueba reciente de esta separación al encontrar que sólo el 55% de 696 informes de investigación proporcionan alguna evidencia sobre la validez de las mediciones.

Junto con la complejidad conceptual, el reto más llamativo que puede contribuir al alejamiento entre la teoría de la validez y la práctica es la incorporación de las consecuencias del uso de los tests como una fuente más de evidencias en el proceso de validación, tal como se propone en la última edición de los estándares (AERA, APA y NCME, 1999). Varias preguntas centran el debate

Fecha recepción: 12-7-05 • Fecha aceptación: 1-12-05
Correspondencia: José Luis Padilla
Facultad de Psicología
Universidad de Granada
18071 Granada (Spain)
E-mail: jpadilla@ugr.es

entre defensores y críticos: el análisis de las consecuencias del uso de los tests y los juicios de valor inherentes a las interpretaciones, ¿deben formar parte de los procesos de validación?, si la respuesta es positiva: ¿qué se debe analizar?, ¿quién es el responsable? y ¿cómo se pueden realizar estos análisis? Para muchos profesionales de la medición, las respuestas a dichas preguntas pueden determinar si la validación se debe considerar una empresa científica, empírica o también un proceso sociopolítico (Crocker, 1997).

Las numerosas propuestas elaboradas para responder a estas preguntas han generado una importante bibliografía. La inclusión del análisis de las consecuencias es examinada en un primer número monográfico de la revista *Educational and Psychological Measurement: Issues and Practice* (vol. 16, nº 4, 1997). Una visión global de las implicaciones de dicha inclusión y, especialmente, del debate filosófico detrás de la misma, puede estudiarse en el número monográfico sobre validez de la revista *Social Indicator Research* (vol. 45, 1998). Un segundo número monográfico de la revista *Educational and Psychological Measurement: Issues and Practice* (vol. 17, nº 2, 1998), aborda desde las perspectivas de los autores de tests, editores, académicos, usuarios y responsables políticos, los retos prácticos que conlleva la aceptación del análisis de las consecuencias sociales como parte del proceso de validación.

Muestra evidente de las posturas enfrentadas puede ser la distancia entre el trabajo de Boorbom, Mellenberg y Heerden (2004), para quienes la «validez del test» no depende de las consecuencias sociales de su uso; frente al de Crocker (2003), al rescatar la noción de «validez aparente» para dar cabida a las expectativas sobre la evaluación de los agentes implicados en los procesos de validación. Además, crece el interés por abordar el análisis de las consecuencias fuera del ámbito de la evaluación educativa. Por ejemplo, un número monográfico de la revista *Human Performance* (vol. 16, nº 3, 2003) se centra en las investigaciones sobre el papel de la «amenaza del estereotipo» para explicar las diferencias entre las medias de grupos minoritarios y mayoritarios en los tests cognitivos utilizados en selección de personal. Gómez y Padilla (2004) mostraron la necesidad de incorporar la evaluación de las consecuencias a las interpretaciones de los tests basadas en estándares de ejecución. A su vez, aumenta el interés de los profesionales encargados de la adaptación de cuestionarios por abordar los aspectos «consecuenciales» de la evaluación transcultural (Casillas y Robbins, 2005; Hambleton, Merenda y Spielberger, 2005).

Tras apuntar los antecedentes históricos y resumir los argumentos más significativos del debate, el objetivo principal del trabajo es ofrecer una respuesta global a la pregunta sobre cuándo la evaluación de las consecuencias afecta a la validez de las inferencias hechas a partir de un test. Se pretende argumentar dicha respuesta a partir de las semejanzas con la evolución de la consideración del sesgo en los tests como un problema de validez. Por último, se discuten los desafíos que la evolución de la noción de justicia en el uso de los tests puede plantear a los estudios de validación de las consecuencias del uso de los tests.

Antecedentes históricos

La exposición del consenso actual y el análisis del debate requieren resumir de forma breve los antecedentes históricos de la preocupación por las consecuencias sociales de la evaluación y del uso de los tests. Resumen orientado a mostrar cómo la evolución de la atención prestada a las consecuencias sociales es difícilmente separable del tratamiento recibido por el problema del sesgo en los tests (Elosua y López-Jáuregui, 2005; Hidalgo, Gómez y Padilla, 2005).

La preocupación por las consecuencias sociales de la evaluación no se inicia en la última década del siglo pasado, como se podría pensar a raíz de su incorporación en los *Estándares* (AERA, APA y NCME, 1999). Aunque centrados en contextos específicos, algunos autores habían abordado con anterioridad las repercusiones del uso de los tests para la selección de personal (Guion, 1974), o en la medición del rendimiento académico (Ebel, 1961). De hecho, se puede recurrir a los estándares anteriores para rastrear los antecedentes:

a) Primeros estándares y recomendaciones técnicas. Los primeros estándares profesionales fueron en realidad dos documentos. El primero, titulado *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, fue publicado en 1954 por la APA. Un año después, la AERA y el NCME editaban el segundo: *Technical Recommendations for Achievement Tests*. Las recomendaciones técnicas de la APA apelaban a la «justicia» en la elección de los grupos normativos con los que se interpretara la ejecución de las personas en el test y a que se prestase atención a los nombres dados a los tests para «… minimizar el riesgo de interpretaciones equivocadas por "compradores" y "sujetos"» (p. 10). A su vez, el documento de la AERA y el NCME recomendaban que si «… podía ser razonable esperar que la validez fuese diferente en subgrupos identificables… el manual debía informar de la validez para cada subgrupo por separado» (p. 26).
b) El término «*Standards*» aparece en la edición de 1966, realizada ya de forma conjunta por las tres asociaciones (AERA, APA y NCME). Se alerta frente a los juicios de valor que pueden desencadenar nombres confusos como tests «libres de cultura», de «habilidades mentales primarias» o de «creatividad» (p. 9). A su vez, en el comentario al Estándar B1.5 exponen como ejemplo de interpretación equivocada de las mediciones los casos de baja ejecución en los tests de aptitudes de personas con poco dominio del idioma del test (p. 10). Al igual que en la edición anterior no aparece el término «sesgo».
c) Los *Estándares* de 1974 (AERA, APA y NCME) marcan una nueva etapa en la preocupación por las consecuencias del uso de los tests. Se reconoce de forma explícita que parte del estímulo para esta nueva edición son las acusaciones de discriminación a los grupos minoritarios recibidas por los tests. Consideran que gran parte de las críticas son ejemplos de «*appropriateness*», término incluido en la definición de validez de los últimos estándares, como: inadecuación del idioma, antecedentes culturales de los examinados, cantidad y calidad de ciertos tipos de entrenamiento, etc. Dentro del capítulo dedicado a la fiabilidad y la validez, el Estándar E9 plantea que «el usuario del test debe investigar la posibilidad de sesgo en el test o en los ítems del test» (p. 43). Asimismo, se recomienda investigar posibles diferencias en la validez relacionada con el criterio para submuestras definidas por variables demográficas. El comentario que acompaña a este estándar es un análisis de las diferentes definiciones de justicia y sus implicaciones para el estudio del sesgo.
d) Los *Estándares* de 1985 (AERA, APA y NCME) suponen una continuidad respecto del tratamiento de las consecuencias

> sociales de la evaluación. El Estándar 6.5 plantea que «Los usuarios de los tests deben estar alerta a consecuencias probables no deseables del uso de los tests e intentar evitar acciones que tengan consecuencias negativas indeseables» (p. 42). Además, en los párrafos dedicados a la «predicción diferencial» analizan el problema del «sesgo en selección» intentando separarlo del problema de la justicia al exponer «A diferencia del sesgo en selección…, la justicia no es un término psicométrico técnico; está sujeto a definiciones diferentes en diferentes circunstancias políticas y sociales» (p. 13).

Esta somera revisión del contenido de las diferentes ediciones de los estándares muestra ya la dificultad de separar los análisis de las consecuencias sociales de la evaluación, de la evolución del problema del sesgo y de las nociones de justicia en el uso de los tests. Paralelismo en el tratamiento de estas cuestiones apuntado también por Cole y Zieky (2001), al establecer tres períodos en la evolución del tema de la justicia con puntos intermedios coincidentes con los *Estándares* de 1974 y la última de 1999. A su vez, Linn (2001) reconoce las ventajas de hacer equivalente la justicia en los tests a una validez comparable para todas las personas y grupos a la hora de buscar evidencias para responder a multitud de cuestiones sobre la justicia en el uso de los tests.

Antes de abordar el debate desencadenado en los años noventa del siglo pasado, previo al consenso actual reflejado en la última edición de los *Estándares* (APA, AERA, NCME, 1999), resulta imprescindible resaltar las aportaciones de Samuel Messick al debate sobre la incorporación al análisis de la validez de las consecuencias sociales y cómo su obra refleja también la interrelación entre dicho análisis, el problema del sesgo y el de la justicia en el uso de los tests. Messick (1975) publica un artículo en la revista *American Psychologist*, con el significativo subtítulo de «Meaning and values in measurement and evaluation», donde afirmaba que los valores impregnan las decisiones sobre el uso de los tests para objetivos específicos. Formula también las dos preguntas que en su opinión se deben responder cuando se plantea el uso de un test para un objetivo concreto: «Primero, ¿es el test suficientemente bueno como medida de la característica que pretende evaluar? Segundo, ¿debería utilizarse el test para el objetivo propuesto?» (p. 962). Considera que la primera pregunta es una cuestión técnica y científica, mientras que la segunda es ética, requiriendo una evaluación de las consecuencias sociales del uso del test en términos de valores sociales. La apuesta por la incorporación de las consecuencias sociales al concepto de validez es tan firme como para considerar la validez un imperativo ético del uso de los tests (Messick, 1980). Además, plantea en este trabajo un paralelismo interesante entre las dos preguntas anteriores y dos críticas recurrentes hacia los tests: (a) la cuestión del sesgo en los tests o la adecuación de la medida, y (b) la cuestión de la justicia de los tests o la pertinencia de su utilización. La referencia al sesgo no es gratuita, ya que le permite proponer aproximaciones para la valoración de las consecuencias del uso de los tests. Pero, sin duda, la aportación más influyente es su capítulo en la tercera edición del *Educational Measurement* (Linn, 1989); capítulo que sustituyó al de Cronbach (1971) como la referencia más citada y de mayor impacto sobre la teoría de la validez (Shepard, 1993). De esta obra procede la definición de validez citada en casi todos los trabajos posteriores y de clara influencia en la recogida una década más tarde por la última edición de los *Estándares*: «un juicio integrado del grado en el que la evidencia empírica y la racionalidad teórica apoyan la suficiencia y conveniencia de las inferencias y acciones basadas en las puntuaciones en los tests u otros modos de evaluación» (p. 5).

El debate sobre el papel de las consecuencias

Dados los antecedentes, la novedad al abordar las consecuencias del uso de los tests es su incorporación como un componente esencial dentro del concepto de validez (Moss, 1992). Crocker (1997) resalta la importancia de esta incorporación en su calidad de editora del número monográfico al que nos hemos referido antes (*Educational Measurement: Issues and Practice, 16 (2)*), al tiempo que expone el núcleo del debate: «La cuestión es si la investigación de las posibles consecuencias de administrar y utilizar una evaluación deberían ser definidas como una parte integral del plan de validación» (p. 4). A continuación, se resumen las posturas a favor y en contra de los autores que contribuyen a este monográfico para cerrar el apartado con la réplica de Messick (1995).

A favor de la incorporación

Según Shepard (1997), la mayoría de los especialistas reconocen la utilidad y necesidad de atender a las cuestiones sobre «justicia social» y «efectos del uso de los tests», pero discrepan sobre si tales cuestiones deberían ser consideradas parte de la validez del test. Asimismo, se queja de que los participantes en el debate han creado la falsa impresión de que se ha introducido un nuevo tipo de validez —la validez consecuencial— y, en consecuencia, demandan una definición más rigurosa de la misma. No obstante, el argumento fuerte de la postura de Shepard (1997) es que las consecuencias pueden formar parte del esquema conceptual donde se incluye el constructo objeto de la medición y que orienta el proceso de validación; de hecho, no serían más que hipótesis rivales de la interpretación prevista de las puntuaciones. Así, dicho esquema conceptual conecta las puntuaciones con las consecuencias, argumentos que deben ser objeto de la «validación consecuencial». Para apoyar este argumento recurre al ejemplo desarrollado por Kane (1992): al analizar la validez de un test de álgebra utilizado como prerrequisito para un curso de cálculo, no se debe sólo probar que el test de álgebra es una medida válida de las destrezas previstas de álgebra, sino que cuando se utiliza para una «clasificación» diferencial, se debe demostrar que los estudiantes con bajas puntuaciones en el test de álgebra harán mejor el curso de cálculo si siguen antes un curso intensivo de álgebra, antes de hacer el curso de cálculo; la relación entre las habilidades prerrequisito de álgebra y el éxito en el curso de cálculo es central en la red nomológica del constructo. Ejemplos evidentes de consecuencias que deben incorporarse al proceso de validación son los que Messick (1989) recoge bajo la etiqueta de «impacto adverso»: identificar a más niños que niñas para clases de educación especial, premiar a más niños que niñas en las escuelas, seleccionar a más blancos que negros para determinados trabajos, etc. Por sí mismas estas consecuencias del uso de los tests no son evidencia de invalidez, pero deben ser objeto de investigación.

Por su parte, para Linn (1997) considerar la validez sólo como la precisión de las inferencias basadas en las puntuaciones transmite una visión demasiado estrecha. Ya la definición de validez de los *Estándares* del 85 (AERA, APA y NCME, 1985) planteaban que «el concepto se refiere a la adecuación, significado y utilidad de las inferencias específicas hechas desde las puntuaciones en el test» (p. 9). Linn (1997) considera que los términos adecuación,

significado y utilidad van más allá del concepto de precisión de las inferencias y conllevan un juicio de valor.

En contra de la incorporación

Popham (1997) defiende la postura contraria a la inclusión de la categoría «validez consecuencial» dentro de la noción de validez que, en su opinión, en este momento es ampliamente entendida como «…la precisión de las inferencias realizadas a partir de la ejecución de los examinados en el test» (p. 9). Según Popham, Messick (1995) va demasiado lejos al abrir la puerta de la validez a las consecuencias sociales cuando plantea que debe ser validada cualquier implicación de la acción que se derive del significado de las puntuaciones, ya que una mala utilización del test no resta validez a la inferencia realizada a partir de él. Tras considerar improductiva la incorporación de la validez consecuencial a la teoría de la validez, apoya su postura en tres argumentos: a) la visión de la validez en los *Estándares* de 1985 es clara y útil, por lo que ahora es el momento de favorecer su aceptación por todos los profesionales de la medición educativa; b) hacer de las consecuencias sociales del uso de los tests un aspecto de la validez introduce una confusión innecesaria en lo que significa la validez de la medición; y c) autores y usuarios de tests deben prestar atención a las consecuencias del uso de los tests, recogiendo las evidencias pertinentes, pero sin que tales evidencias sean una faceta de la validez.

En la misma línea, Mehrens (1997) considera un reduccionismo la idea de que toda validez es validez de constructo y que toda evidencia lo es para o en contra de la validez de constructo. Además, hacer depender la validez de una evaluación de las consecuencias es confundir los resultados de utilizar los datos en un proceso de toma de decisiones —a lo que él hace equivalente la validez consecuencial—, con la precisión de la inferencia sobre la cantidad del atributo medido que posee el individuo. En su opinión se pueden realizar inferencias generales independientes de cualquier uso específico del test.

La réplica de Messick

La aportación de Messick (1998) al número monográfico sobre validez de la revista *Social Indicators Research* (vol. 45), con el significativo título «Test validity: a matter of consequence», contiene un apartado con sus respuestas a los críticos de la incorporación de las consecuencias dentro de la validez de constructo.

Messick (1998) empieza reconociendo las dudas que genera el papel de las consecuencias como fuente de evidencia sobre el significado de las puntuaciones. No obstante, considera que parte del debate se debe a interpretaciones erróneas de sus propuestas. Así, algunos críticos consideran que las consecuencias de errores procedimentales o de interpretaciones equivocadas de las puntuaciones no restan validez a los usos legítimos de los tests. Por el contrario, Messick (1998) explicita que su intención es dirigir la atención hacia los efectos colaterales no anticipados de los usos *legítimos* del test, especialmente si se pueden relacionar los efectos adversos imprevistos con amenazas a la validez del test tales como la baja representación del constructo o fuentes de varianza no relacionadas con el constructo. Las consecuencias imprevistas, al igual que las previstas, forman parte de la red nomológica del constructo y, por tanto, aportan argumentos para analizar el significado de las puntuaciones. Por el contrario, las consecuencias de los malos usos de los tests son irrelevantes para la red nomológica, para el significado de las puntuaciones y para el proceso de validación.

Consecuencias prácticas

El debate avanza y se amplia a cómo se deben examinar las consecuencias, qué obstáculos se plantean a dicho examen y quién tiene la responsabilidad de realizarlo. Diversas contribuciones en el segundo monográfico de la revista *Educational Measurement Issues and Practice* (vol. 17, nº 2, 1998) intentan responder a estas preguntas. El monográfico expone las perspectivas de los distintos agentes implicados en los programas de evaluación: constructores de tests, editores y responsables políticos, y cómo se deben asignar responsabilidades. En general, todas ellas muestran las dificultades del análisis de las consecuencias sociales dentro del proceso de validez. Por ejemplo, Green (1998) considera que la mayoría de los editores no están en una posición que les permita obtener ningún tipo de evidencia sobre las consecuencias del uso de los tests, entre otras razones por: a) la variedad de usos del test que realizan a lo largo del tiempo profesores, escuelas, distritos y estados; b) las escasas posibilidades de que los editores convenzan a todos esos usuarios sobre la descripción más adecuada del constructo; y c) la ausencia de mecanismos directos para obtener evidencias creíbles sobre las consecuencias de los usos de los tests de rendimiento. A pesar de todas las limitaciones expresadas, todos los autores abogan por un esfuerzo coordinado y prolongado en el tiempo de profesionales, autores, editores y asociaciones, para una valoración sistemática de los aspectos consecuenciales de la validez.

¿Cuándo analizar las consecuencias?

Los participantes en la polémica realizaron llamamientos a favor y en contra de que la próxima edición de los estándares reflejara su postura (Linn, 1997; Moss, 1992; Popham, 1997). La respuesta de las tres asociaciones promotoras fue incluir en la última edición disponible de los Estándares (AERA, APA, NCME, 1999) las consecuencias sociales del uso de los tests como una fuente de evidencia más dentro del proceso de validación. El papel atribuido a las consecuencias del uso de los tests como evidencias útiles en el proceso de validación queda claro en la siguiente declaración: «La evidencia sobre las consecuencias puede informar las decisiones sobre la validez. Aquí, sin embargo, es importante diferenciar entre la evidencia que es directamente relevante para la validez y la evidencia que puede informar las decisiones sobre política social pero que cae fuera del alcance de la validez» (p. 16).

Los *Estándares* de 1999 ofrecen la respuesta consensuada en el ámbito profesional a la pregunta ¿Cuándo incorporar el análisis de las consecuencias del uso de los tests durante el proceso de validación? Siempre que la validez de la interpretación deseada de las mediciones pueda resultar amenazada por evidencias de «baja representación del constructo» o por la presencia de fuentes de variación «irrelevantes al constructo». El test adolece de baja representación cuando los ítems no representan de forma adecuada todos los componentes importantes del constructo; mientras que pueden aparecer fuentes de varianza irrelevantes cuando factores extraños al constructo objeto de la medición afectan a las puntuaciones.

Al tratar de ejemplificar estas situaciones, los redactores de los estándares recurren como ejemplo a la aparición de diferencias significativas en las tasas de contratación entre miembros de dife-

rentes grupos demográficos como resultado de las mediciones aportadas por un test. Así, el hallazgo de diferencias grupales no amenaza la validez de la interpretación deseada cuando se deban solamente a una distribución diferente de las habilidades que persigue medir el test. Sin embargo, si el test mide diferencias en capacidades no relacionadas con la ejecución en el trabajo —«baja representación del constructo»—, o si las diferencias son debidas a la sensibilidad del test a algunas características de los sujetos que no se deseaba fueran parte del constructo —«fuentes irrelevantes de invarianza»—, entonces la validez debe ser puesta en cuestión, incluso si las puntuaciones en el test correlacionan positivamente con alguna medida de la ejecución en el trabajo. Intentando precisar algo más el ejemplo, ante el hallazgo de las diferencias en las tasas de contratación, se trataría de obtener evidencias sobre la hipótesis rival de que las diferencias se deban a que los grupos tienen un dominio promedio diferente de componentes irrelevantes del constructo, o a un dominio menor del idioma del test en un grupo frente al otro. Se trataría, por tanto, de incorporar proposiciones sobre el origen de las diferencias grupales contrarias a la interpretación deseada de las mediciones, por ejemplo: diferentes estrategias a la hora de responder a los tests; familiaridad diferencial con formatos de ítems y contenidos, tipos particulares de entrenamiento para responder, etc.

En principio, puede llamar la atención el reducido número de estándares (E. 1.23 y E 1. 24) dedicados a la evaluación de las consecuencias dentro del capítulo de validez; sin embargo, el paralelismo con el tratamiento del sesgo y el de la justicia en el uso de los tests vuelve a quedar patente al constatar que la evaluación de las consecuencias es tratada en la Parte II de los Estándares, bajo el título general de «Fairness in testing», y, en concreto, los aspectos más directos de la relación entre la evaluación de las consecuencias y el sesgo en el capítulo 7, titulado «Fairness in testing and test use».

Conclusiones

La inclusión de la evaluación de las consecuencias sociales del uso de los tests en la última edición de los *Estándares* (AERA, APA y NCME, 1999), como una fuente más de evidencia, es la respuesta consensuada entre los profesionales a la preocupación histórica y al debate de los años noventa del siglo pasado sobre las consecuencias sociales de la evaluación. Centrar las evidencias en la «baja representación del constructo» y la presencia de «fuentes irrelevantes de varianza» para decidir cuándo las consecuencias sociales amenazan la validez, refleja el permanente deseo de los profesionales por no convertir la validación en un proceso sociopolítico (Crocker, 1997). Postura bien acogida por las semejanzas evidentes con el tratamiento histórico del sesgo en los tests. Prueba de esta semejanza es la definición del término «sesgo en el test» recogida en los *Estándares actuales*: «… se refiere a componentes irrelevantes para el constructo que originan sistemáticamente puntuaciones más altas o más bajas para grupos identificables de examinados. Tales componentes pueden ser introducidos por un muestreo inapropiado del contenido del test» (p. 77). Esta semejanza es fácilmente detectable en los estándares dedicados al sesgo y al DIF en el capítulo sobre la justicia en el uso de los tests.

Sin embargo, no parece fácil que la evaluación de las consecuencias del uso de los tests se mantenga en los límites fijados por los últimos *Estándares*. Cole y Zieky (2001) reconocen el avance que ha supuesto el tratamiento de la justicia en el uso de los tests en los *Estándares*, pero a la vez muestran su insatisfacción por haber unido el problema de la justicia al de la validez por medio de la evaluación de las consecuencias sociales. Esta «unión» impide hacer juicios inequívocos sobre la justicia de un test, dado que la validez es siempre una cuestión de grado. Sin contar con las diferencias en las nociones de justicia debidas en gran parte a valores diferentes para todas las personas implicadas en un proceso de evaluación. El «programa» de futuro propuesto para el estudio de la justicia en el uso de los tests: disminuir las diferencias grupales no relacionadas con el constructo objeto de la medición, diseñar oportunidades equiparables de ejecución en los tests, como reflejan los capítulos 9 y 10 de los últimos *Estándares* dedicados al uso de los tests en personas con diferentes antecedentes lingüísticos y discapacidades, desterrar los malos usos y adaptar los tests a las diferencias individuales; generará importantes desafíos no sólo a los procesos de validación, sino a la propia elaboración de instrumentos de evaluación (Willingham, 2001).

En definitiva, puede decirse que hay un amplio consenso en la incorporación de las consecuencias al proceso de validación de la interpretación deseada de las mediciones. Pero también que el debate continuará enriqueciéndose y que se necesitan metodologías que den respuesta a la búsqueda de evidencias sobre las consecuencias del uso de los tests. Además, resta por ver si el análisis de las consecuencias se mantendrá en los límites actuales o la presión social, que tarde o temprano se extiende a todos los países dentro de un mismo marco cultural, hará que se deban contemplar los efectos adversos del uso de los tests, estén o no ligados a fuentes de invalidez.

Nota

Este trabajo ha sido financiado por el Ministerio de Educación y Ciencia: Proyectos SEJ2005-09144- C02-02 y SEJ2005-08924-PSIC.

Referencias

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

American Educational Research Association y National Council on Measurement in Education (1955). *Technical recommendations for achievement tests*. Washington, DC: National Education Association.

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

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

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

Borsboom, D., Mellenbergh, G.J. y J. van Heerden (2004). The concept of validity. *Psychological Review, 111*, 1.061-1.071.

Casillas, A. y Robbins, S.B. (2005). Test adaptation and cross-cultural assessment from a business perspective: issues and recommendations. *International Journal of Testing, 5*, 5-21.

Cole, N.S. y Zieky, M.J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369-382.

Crocker, L. (1997). Editorial: the great validity debate. *Educational Measurement: Issues and Practice, 16*, 4.

Crocker, L. (2003). Teaching for the test: validity, fairness and moral action. *Educational Measurement: Issues and Practice, 22*, 5-11.

Cronbach, L.J. (1971). Test validation. En R.L. Thorndike (ed.): *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Ebel, R.L. (1961). Must all test be valid? *American Psychologist, 16*, 640-647.

Elosua, P. (2003). Sobre la validez de los tests. *Psicothema, 15*, 315-321.

Elosua, P. y López-Jáuregui, A. (2005). Clases latentes y funcionamiento diferencial del ítem. *Psicothema, 17*, 516-521.

Gómez, J. y Padilla, J.L. (2004). The evaluation of consequences in standard-based test scores interpretations. *Measurement, 2*, 104-108.

Green, D.R. (1998). Consequential aspects of the validity of achievement tests: a publisher's point of view. *Educational Measurement: Issues and Practice, 17*, 16-20.

Guion, R.M. (1974). Open a new window: validities and values in psychological measurement. *American Psychologist, 29*, 287-296.

Hambleton, R.K., Merenda, P.F. y Spielberger, C.D. (eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment.* London: LEA.

Hambleton, R.K. y Pitoniak, M.J. (2002). Testing and measurement. En J. Wixted (ed.): *Methodology in experimental psychology* (pp. 517-562). En H. Pashler (ed.): Stevens' Handbook of Experimental Psychology (third edition). New York: John Wiley & Sons, Inc.

Hidalgo, M.D., Gómez, J. y Padilla, J.L. (2005). Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial de los ítems. *Psicothema, 17*, 509-515.

Hogan, T.P. y Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*, 802-812.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

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

Linn, R. (2001). Constructs and values in standards-based assessment. En H.I. Braun, D.N. Jakson y D.Wiley (eds.): *The role of constructs in psychological and educational measurement* (pp. 231-254). New Jersey: Lawrence Erlbaum Associates, Publishers.

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

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*, 1.012-1.027.

Messick, S. (1989). Validity. En R.L. Linn (ed.): *Educational measurement* (pp. 13-103). New York: MacMillan.

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

Messick, S. (1998). Test validity: a matter of consequence. *Social Indicators Research, 45*, 35-44.

Moss, P.A. (l992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Muñiz, J. (2003). La validación de los tests. *Metodología de las Ciencias del Comportamiento, 5*, 119-139.

Popham, W. (1997). Consequential validity: right concern-wrong concept. *Educational Measurement: Issues and Practice, 6, 9-14.*

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

Shepard, L.A. (1997). Evaluating centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 6*, 71-86.

Willingham, W.W. (2001). Seeking fair alternatives in construct design. En H.I. Braun, D.N. Jakson y D. Wiley (eds.): *The role of constructs in psychological and educational measurement* (pp. 193-206). New Jersey: Lawrence Erlbaum Associates, Publishers.