Psicothema 2007. Vol. 19, nº 1, pp. 173-178
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los test

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

El esquema de validación basado en argumentos orienta la evaluación de las consecuencias del uso de los tests. La distinción entre «inferencias semánticas» e «inferencias políticas» permite integrar la validación de las consecuencias en un esquema único de validación. El proceso de validación debe aportar evidencias sobre los supuestos que sostienen ambos tipos de inferencias. Tras presentar el esquema de validación, se ejemplifica su utilización a través de la evaluación del uso de los tests en dos aplicaciones: el uso de tests de alto riesgo en el contexto educativo y la validación de las adaptaciones para personas con discapacidades en los tests estandarizados. Por último, se proponen procedimientos para la validación de las consecuencias y se discute la relevancia del esquema de validación basado en argumentos para la validación de las consecuencias del uso de los tests en el contexto español.

Validation scheme and procedures to analyze consequential validity. The argument-based validation scheme guides assessment of the consequences of testing. The distinction between «semantic inference» and «political inference» allows us to combine the validation of the consequences in a single validation scheme. The validation process should produce evidence about the assumptions that support both types of inference. After presenting the validation scheme, we provide examples of its use in the assessment of the testing of two applications: the use of high-stake tests in educational contexts and the validation of adjustments made in standardized tests for people with disabilities. Finally, we propose procedures for the validation of consequences and we discuss the suitability of the argument-based validation scheme for the validation of the consequences of testing in Spain.

La evaluación de las consecuencias del uso de los tests en los procesos de validación es uno de los mayores desafíos planteados por la versión actual de la teoría de la validez (Messick, 1998). Desafío conceptual y metodológico, al situar la construcción y uso de los tests en un escenario donde resulta difícil diferenciar entre las cuestiones de validez y los argumentos ideológicos, políticos o sobre la justicia en el uso de los tests (Crocker, 1997).

Antes y después de que la última edición de los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1999) incluyera la validación de las consecuencias como una fuente más de evidencia, se han expuesto las dificultades de su incorporación a la teoría de la validez. Dificultades agrupables en torno a dos preguntas: ¿qué evaluar? y ¿cómo hacerlo? Para dar respuesta a la primera, Messick (1998) abogó por dirigir la mirada hacia los efectos imprevistos del uso de los tests, sobre todo cuando se pudieran relacionar con amenazas a la validez procedentes de la baja representación del constructo o de fuentes de variación no relacionadas con el constructo. Ésta fue la recomendación seguida por los editores de los *Standards*. Así, el Estándar 1.24 recomienda: «Cuando resulten consecuencias imprevistas del uso de los tests, se debe realizar un intento por investigar si tales consecuencias proceden de la sensibilidad del test a características distintas de las que estaba previsto evaluar o de fallos del test al representar completamente el constructo previsto» (p. 23). Padilla, Gómez, Hidalgo y Muñiz (2006) mostraron cómo la semejanza entre la respuesta dada por los *Standards* a la pregunta sobre el contenido de la validación de las consecuencias y la definición de sesgo había facilitado la incorporación de las primeras al consenso actual en torno a la teoría de la validez. Sin embargo, las dudas acerca de qué validar permanecen por la ausencia en los *Standards* de indicaciones claras sobre cuándo un test falla al representar un constructo o cómo identificar fuentes de varianza ajenas al constructo.

A la falta de indicaciones claras hay que sumar la amplitud y dificultad de las cuestiones a investigar. Reckase (1998) alertó de que la base consecuencial de la validez incluía consideraciones difíciles de atender durante la elaboración de un test. Por un lado, la «base consecuencial para la interpretación del test» obliga a examinar, primero, los juicios de valor inherentes a la denominación del constructo y a la teoría sobre el constructo; y, segundo, el marco ideológico donde se encuadra dicha teoría. Por otro lado, la «base consecuencial sobre la utilización del test» requiere anticipar las consecuencias potenciales o reales de aplicar el test (Messick, 1998). Si no resulta fácil trabajar con valores y perspectivas ideológicas, ¿cómo establecer relaciones causales entre el uso de un test en un programa de evaluación y todas las posibles consecuencias de esa evaluación para las partes implicadas? Muñiz

Fecha recepción: 5-6-06 • Fecha aceptación: 19-9-06
Correspondencia: José Luis Padilla
Facultad de Psicología
Universidad de Granada
18071 Granada (Spain)
E-mail: jpadilla@ugr.es

(2003) señaló el peligro de que la complejidad de estas cuestiones distrajera a los constructores de tests de su misión principal: aportar evidencias sobre la representación del constructo conseguida en el test.

La cuestión de cómo evaluar las consecuencias tampoco tiene una respuesta fácil. Por un lado, a diferencia de las otras fuentes de evidencias que se benefician de la existencia de técnicas y procedimientos tradicionalmente utilizados en los estudios de validación, no hay mecanismos directos para obtener evidencias creíbles sobre las consecuencias del uso de los tests (Green, 1998). Por otro, la semejanza conceptual entre la validación de las consecuencias y la definición de sesgo que permite utilizar todo el arsenal metodológico de los estudios de sesgo, no parece suficiente para tratar con la variedad y complejidad de las cuestiones que reclaman la atención de los profesionales de la evaluación psicológica y educativa. Pruebas de la necesidad de nuevos enfoques metodológicos son el debate generado por la validación de las consecuencias de los denominados «tests de alto riesgo» dentro de los programas americanos de evaluación educativa (Camilli, 2003), la investigación sobre las repercusiones para la validez de las acomodaciones o adaptaciones de los tests para personas con discapacidades (Sireci, 2004), los efectos de las atribuciones y juicios de valor sobre la ejecución en los tests de aptitudes (Nguyen, O'Neal y Ryan, 2003), o las apelaciones a incorporar evaluaciones sobre la «sensibilidad cultural» y la «justicia» en las adaptaciones de tests a diferentes grupos lingüísticos y culturales (Casillas y Robbins, 2005).

Las dificultades apuntadas, junto con la variedad de exigencias que se pueden plantear para justificar el uso de los tests en los diferentes contextos de evaluación (educativo, clínico, selección laboral, etc.), hacen necesario abordar la validación de las consecuencias desde un esquema conceptual que oriente el proceso de validación (Moss, 2003). Además, dentro del esquema conceptual es donde cobran sentido el recurso a nuevas aproximaciones metodológicas con las que obtener evidencias creíbles sobre el efecto de las consecuencias en la validez de las mediciones. La validación basada en argumentos de validez (Cronbach, 1988; Kane, 1992, 2001, 2002) puede ser el esquema conceptual idóneo desde el que abordar la validación de las consecuencias del uso de los tests.

El principal objetivo de este trabajo es mostrar la adecuación del esquema de validación basado en argumentos de validez para obtener evidencias sobre las consecuencias del uso de los tests. Tras apuntar los contenidos del esquema conceptual, se ejemplifica su utilización revisando investigaciones recientes sobre la validación de las consecuencias del uso de tests de «alto riesgo» en el contexto de la evaluación educativa y sobre la validación de las adaptaciones en los tests estandarizados para personas con discapacidades. Por último, se apuntan propuestas metodológicas y se muestra la relevancia del esquema conceptual para la validación de las consecuencias en el contexto cultural español.

Aproximación a la validación basada en argumentos

Abordar la validación desde un esquema conceptual basado en argumentos de validez es una recomendación recogida en los propios *Standards*: «La validación puede ser entendida como la elaboración de un argumento de validez científicamente sólido que apoye la interpretación prevista de las puntuaciones en el test y su relevancia para el uso propuesto» (AERA, APA, y NCME, p. 9). Cronbach (1988), tras proponer extender a los tests todas las lecciones aprendidas de la metodología de evaluación de programas, animó a los profesionales para que pensaran en la validación más en términos de elaboración de un «argumento de validez» que en los de una «investigación de validez».

Resulta necesario para valorar el alcance de la aproximación a la validación basada en argumentos, identificar la connotación más pertinente en este contexto de la palabra «argumento». Puede ser adecuado atribuirle el significado de «alegato» casi en un sentido jurídico, en cuanto que se trata de un juicio o evaluación global sobre la plausibilidad de la interpretación propuesta de las mediciones y de la justificación del uso previsto del test. El carácter «convincente» que debe tener el argumento queda justificado al dirigirse la validación a una audiencia diversa y potencialmente crítica ante el uso del test; de ahí que el argumento deba «conectar conceptos, evidencias, consecuencias personales y sociales, y valores» (Cronbach, 1988, p. 4).

La presentación más elaborada de la aproximación a la validación basada en argumentos se encuentra en los trabajos de Michael T. Kane (Kane, 1992, 2001, 2002). Elaborar el argumento de validez requiere aclarar el contenido de la interpretación propuesta de las mediciones. Kane (2002) sugiere que el camino para realizar esta aclaración es especificar un «argumento interpretativo», es decir, «… el esquema de inferencias y supuestos que conducen desde las puntuaciones a las conclusiones y decisiones» (p. 31). A través de un ejemplo sobre la interpretación «deseada» de las puntuaciones en un test de rendimiento en matemáticas utilizado como medida de la capacidad de solución de problemas, Kane (2002) establece cuatro pasos generales para el desarrollo de un argumento interpretativo aplicables a muchos procesos de evaluación educativa: 1) «Evaluación» de la ejecución del estudiante en cada ítem dando lugar a la asignación de una puntuación y a la combinación de las puntuaciones individuales en una única puntuación observada para cada estudiante; 2) «Generalización» desde la ejecución realmente observada a conclusiones sobre la ejecución esperada en un universo de problemas similares en condiciones semejantes; 3) «Extrapolación», extendiendo las conclusiones al dominio de problemas de matemáticas indicadores de la capacidad de solución de problemas; y 4) «Decisión», sobre el futuro del examinado (e. g., graduación, admisión, certificación, etc.), a partir de las puntuaciones obtenidas en el test. Cada una de estas inferencias generales y los supuestos en que se basan (e. g., precisión de las mediciones, ausencia de fuentes sistemáticas de error, etc.) deben contar con las evidencias adecuadas para que el argumento interpretativo en su conjunto sea considerado válido.

La formulación rigurosa de un argumento interpretativo permite aclarar el contenido de la interpretación propuesta de las mediciones al tiempo que proporciona un esquema para desarrollar el argumento de validez. El argumento interpretativo orienta la validación al señalar el tipo de evidencia de validez más necesario (Kane, 1992). Dada la dependencia entre las inferencias, el foco de la validación deben ser las partes más débiles del argumento interpretativo (Kane, 2001).

La validación de las consecuencias en la aproximación basada en argumentos

Obtener las evidencias adecuadas para validar las consecuencias del uso de los tests no es una tarea sencilla. Primero, porque las mismas puntuaciones en el test pueden utilizarse para tomar una amplia variedad de decisiones. Así, en el ejemplo propuesto

por Kane (2002), las puntuaciones de los examinados en el test de rendimiento en matemáticas podrían servir para decidir sobre su graduación, su admisión en un centro universitario, para valorar la efectividad del programa instruccional, como base de la «rendición de cuentas» del centro educativo, etc. A esta variedad de posibles decisiones se suman las diferencias entre los enfoques con los que los agentes legítimamente implicados en el uso del test (examinados, familias, profesores, responsables académicos, asociaciones, etc.) valorarán el proceso de evaluación juzgando las consecuencias desde diferentes perspectivas, intereses y valores.

Con el objeto de incorporar la evaluación de las consecuencias al esquema de validación, Kane (2001) propuso diferenciar entre la «parte descriptiva» y la «parte prescriptiva» del argumento interpretativo. Dentro de la primera, quedarían incluidas las «interpretaciones descriptivas»: aquellas que estiman alguna variable de los examinados sin especificar ningún uso concreto para las puntuaciones; mientras que en la segunda se incluirían las «interpretaciones basadas en la decisión». Estas últimas implican supuestos que apoyan la adecuación del procedimiento de decisión dentro de un plan de acción o de una decisión política. Según Kane (2002), la justificación de los planes de acción o de las decisiones políticas se hace con referencia a sus consecuencias, es decir, a los resultados «deseables» o «indeseables» de las mismas: «La clave está en los juicios de valor (¿qué es deseable?), y en los supuestos empíricos (sobre los resultados esperados) implicados en la justificación del uso del test» (p. 32).

La separación entre los componentes descriptivo y prescriptivo del argumento pretende aclarar el alcance del proceso de validación, al tiempo que incorpora la evaluación de las consecuencias del uso de los tests dentro de un único esquema de validación. Los propios *Standards* reconocen el papel de la validación tanto para ocuparse de las «interpretaciones descriptivas» como de las «interpretaciones basadas en la decisión»: «La validación comienza con una declaración explícita de la interpretación propuesta para las puntuaciones en el test, junto con una justificación de la relevancia de la puntuación para el uso propuesto» (AERA, APA y NCME, 1999, p. 9). De hecho, las interpretaciones basadas en la decisión habitualmente implican un componente descriptivo como parte del argumento interpretativo. Por ejemplo, las decisiones sobre el «futuro» del examinado (e. g., graduación, admisión, etc.), a partir de sus puntuaciones en el test de rendimiento en matemáticas partirían de inferencias sobre su nivel esperado de ejecución en el universo de problemas semejantes a los del test o en el dominio de problemas de matemáticas que requieran habilidades de solución de problemas.

Se completa la inclusión de la evaluación de las consecuencias dentro del esquema de validación basado en argumentos, diferenciando entre los «supuestos semánticos» que sustentan las interpretaciones descriptivas al dotar de significado las puntuaciones en los tests; y los «supuestos políticos» que apoyan las interpretaciones basadas en la decisión, denominadas también «inferencias políticas», con referencia a sus consecuencias. De forma general, las evidencias que sustentan los supuestos políticos pretenden justificar el uso del test mostrando que las consecuencias positivas superan a las potenciales consecuencias negativas.

A su vez, el esquema de validación basado en argumentos permite validar procesos de evaluación donde las consecuencias pueden afectar al significado de las mediciones bien a través de la aparición de fuentes de variación no relacionadas con el constructo o de una mala representación del constructo por el test. Ejemplos claros de esta «interacción» entre la parte descriptiva y prescriptiva del argumento interpretativo son las consecuencias de la evaluación educativa del «alto riesgo» (e. g., Haladyna y Downing, 2004), o el efecto de las acomodaciones en los tests estandarizados para las personas con discapacidades (Sireci, 2006).

Validación de las consecuencias del uso de los tests en la evaluación educativa

En los últimos años se ha generado en Estados Unidos un intenso debate sobre las consecuencias del uso de los tests en los procesos de evaluación educativa de «alto riesgo» (Camilli, 2003; Cizek, 2001; Mehrens, 1997). Los *Standards* denominan «tests de alto riesgo» a los tests que proporcionan mediciones con consecuencias directas importantes para los examinados, los programas o las instituciones implicadas en la evaluación (AERA, APA y NCME, 1999). Haladyna y Downing (2004) apuntan entre las decisiones de «alto riesgo»: la promoción y graduación de los estudiantes; la evaluación de los centros educativos, acompañada de un sistema de sanciones y recompensas en función de sus logros; o las repercusiones sobre el empleo de los profesores en función de los resultados de sus alumnos en los tests. Ayuda a valorar la importancia del debate sobre el uso de los tests de «alto riesgo», conocer que los programas políticos de reforma educativa en Estados Unidos consideran a los tests estandarizados instrumentos imprescindibles para elevar el rendimiento educativo y para movilizar a todos los agentes implicados en la consecución de los objetivos de las reformas (Kane, 2002).

Numerosos trabajos han abordado la evaluación de las consecuencias del uso de los tests de alto riesgo utilizando el esquema de validación basado en argumentos de validez (e.g., Haladyna y Downing, 2004; Kane, 2002; Lane y Stone, 2002). Estos trabajos no sólo ilustran la potencialidad del esquema de validación para la validación de las «interpretaciones basadas en la decisión» o «inferencias políticas», sino que además han abierto el camino para estudiar los casos en los que las consecuencias interactúan con las «interpretaciones descriptivas» o «interpretaciones semánticas», modificando el significado del constructo.

Kane (2002) ejemplifica la elaboración de un argumento interpretativo para la evaluación de las consecuencias del uso de los tests de graduación. Incluye en la parte descriptiva del argumento tres inferencias semánticas generales: 1) el rendimiento en el test es utilizado para estimar el rendimiento en los «estándares del test» (i. e., el subconjunto de los objetivos educativos fijados para esa etapa educativa incluidos en las especificaciones del test); 2) el rendimiento inferido en los estándares del test es utilizado para estimar el rendimiento en el conjunto de los estándares; y 3) el rendimiento estimado en el conjunto de los estándares es empleado para estimar el rendimiento global al finalizar la enseñanza. Estas tres inferencias generales, junto con los supuestos semánticos en que se apoyan, dotan de significado a las puntuaciones en el test. A continuación, formula una sola inferencia política dentro del argumento interpretativo: los estudiantes con puntuaciones en el test por encima de una puntuación de corte establecida son premiados con el diploma de graduación y aquellos con puntuaciones inferiores no son premiados. Los dos supuestos políticos que sustentan la inferencia muestran cuáles deben ser los objetivos de los estudios de validación sobre las consecuencias del uso de los tests: a) la utilización del test de graduación elevará los niveles de rendimiento sobre los estándares educativos; y b) la utilización del

test de graduación no tendrá un impacto negativo sobre el rendimiento en áreas no medidas con el test.

Lane y Stone (2002) abogan por una evaluación más comprehensiva de las potenciales consecuencias negativas a través de la participación de múltiples agentes implicados (examinados, profesores, responsables académicos, etc.), en la elaboración del argumento interpretativo. Así reclaman estudios que aporten evidencias sobre potenciales consecuencias negativas del uso de los tests de alto riesgo, entre las que citan como candidatas: a) el «estrechamiento» del currículo y focalización de la instrucción sólo en los estándares evaluables con el test; b) el uso «no ético» de materiales y estrategias instruccionales para entrenar en responder al test; c) la ejecución diferencial de subgrupos de estudiantes con diferentes oportunidades de aprender, etc.

Varios de estos potenciales efectos negativos están incluidos en la taxonomía de fuentes de varianza no relacionadas con el constructo, elaborada por Haladyna y Downing (2004). La taxonomía agrupa dichas fuentes de varianza en cuatro grandes áreas: a) uniformidad y tipo de preparación para el test; b) puntuación, aplicación y elaboración del test; c) características diferenciales de los estudiantes; y d) fraude individual e institucional. Sin duda, muchas de las fuentes de varianza incluidas en la taxonomía prueban la «interacción» entre consecuencias derivadas del uso previsto para los tests y amenazas a la interpretación deseada de las mediciones.

Consecuencias de las adaptaciones en los tests para personas con discapacidades

El estudio de las adaptaciones en los tests estandarizados para personas con discapacidades ha aumentado en los últimos años (Sireci, Li, y Scarpati, 2003). El fin último de las adaptaciones es incrementar la validez de las mediciones de los conocimientos y habilidades de estas personas, eliminando las barreras que para ellas puede conllevar la estandarización (e. g., presentando una versión en Braille del test para personas con deficiencias visuales, alargando el tiempo previsto para la realización del examen, etc.). Al incrementar la validez se persigue promover la igualdad de oportunidades educativas y laborales de colectivos tradicionalmente desfavorecidos. Sireci (2005) ha sintetizado en dos preguntas generales el desafío que plantean las adaptaciones a los profesionales de la medida: «¿Las puntuaciones obtenidas de administraciones no estandarizadas del test tienen el mismo significado que las que resultan de administraciones estandarizadas? ¿Conducen las adaptaciones en el test a interpretaciones más válidas para ciertos grupos de estudiantes?» (p. 3). Responder a estas preguntas lleva a enfrentarse a lo que el propio autor califica de oxímoron psicométrico: adaptar un test estandarizado.

Numerosas directrices de asociaciones profesionales recomiendan prestar atención a la cuestión de las adaptaciones en los tests. En el marco europeo los *European Test User Standards for Test Use in Work and Organizational Setting* incluyen el Estándar 2.3 que bajo la etiqueta «Dar la consideración debida a las cuestiones de justicia en los tests», insta en el apartado e) a «realizar las adaptaciones pertinentes para las personas con discapacidades que respondan a los tests» (EFPA y EAWOP, 2005, p. 32). A su vez, la última edición de los *Standards* dedica el capítulo 10 a la evaluación mediante tests de personas con discapacidades. El Estándar 10.1 hace un mandato explícito a constructores, administradores y usuarios para que «…aseguren que la inferencia sobre las puntuaciones en los tests reflejen con precisión el constructo previsto en lugar de cualquier discapacidad y sus características asociadas extrañas al objetivo de la medida» (AERA, APA y NCME, 1999, p. 106).

El esquema de validación basado en argumentos permite responder de forma ordenada a las cuestiones de validez que plantean las adaptaciones. Retomando la terminología propuesta por Kane (2002), el problema arranca de una «inferencia general política»: las adaptaciones igualan las oportunidades de éxito de las personas con y sin discapacidades en las decisiones tomadas a partir de las puntuaciones en el test. La validez de esta inferencia descansa en una inferencia descriptiva: las adaptaciones no modifican el significado del constructo para las personas que responden a la versión adaptada o no del test.

Sireci (2006) formula la hipótesis general contraria a la interpretación descriptiva anterior al plantear que las adaptaciones pueden cuestionar la validez por dos vías: a) la baja representación del constructo (e. g., la lectura en voz alta de los ítems de un test de comprensión lectora para personas con deficiencias visuales); y b) la introducción de fuentes de varianza no relacionadas con el constructo si se aplica el test sin realizar las adaptaciones necesarias (e. g., una persona con deficiencias motoras que no pueda manipular el material del test). En cualquier caso, Sireci (2004) resume la cuestión psicométrica fundamental recurriendo a un concepto familiar en el marco de los procesos de adaptación de tests a través de diferentes grupos lingüísticos y culturales: la validez de las adaptaciones dependerá del grado de «equivalencia del constructo» entre los grupos de examinados que respondan a las diferentes versiones del test.

Hay disponibles revisiones exhaustivas del amplio y variado conjunto de cuestiones planteadas en torno a las adaptaciones de tests estandarizados. Pitoniak y Royer (2001) revisan las cuestiones psicométricas, legales y políticas. Sireci, Li y Scarlati (2003) resumen los resultados de los estudios realizados sobre el efecto de las adaptaciones habituales en las propiedades psicométricas de los tests estandarizados más utilizados. Sireci (2005) aborda una cuestión inevitable cuando no se dispone de evidencia convincente sobre el efecto en la validez de las adaptaciones realizadas: el «etiquetado» o identificación de las puntuaciones obtenidas por personas con discapacidades en los tests adaptados. La cuestión del «etiquetado» genera otro conjunto de efectos consecuenciales que se suman a la necesidad de ofrecer nuevas aproximaciones metodológicas para realizar la validación de las consecuencias del uso de los tests.

Propuestas metodológicas para la evaluación de las consecuencias

La evaluación de las consecuencias del uso de los tests dentro del esquema de validación basado en argumentos implica obtener evidencias sobre la corrección y adecuación de los supuestos en que se basan las inferencias políticas. Al igual que ocurre con la validación de los supuestos que apoyan las inferencias descriptivas, puede recurrirse a cualesquiera de los procedimientos habituales de validación para valorar si las consecuencias positivas del uso del test superan a las potenciales consecuencias negativas.

Sin embargo, la evaluación de las consecuencias presenta una peculiaridad con evidentes repercusiones metodológicas: la necesidad de incorporar al mayor número de partes implicadas en la construcción del argumento interpretativo. No es posible ni deseable ignorar que las diferentes partes interesadas abordarán la validación desde perspectivas, valores e intereses muy variados (Cronbach, 1988). Dar respuesta a esta necesidad lleva a recurrir a una

gran variedad de métodos. Lane y Stone (2002) enumeran entre los métodos utilizados para obtener evidencias sobre los efectos del uso de los tests educativos de alto riesgo: encuestas, entrevistas, grupos de discusión, observaciones en clase, etc.

Cabe resaltar las posibilidades que ofrecen procedimientos como los grupos de discusión y las entrevistas para obtener evidencias sobre las perspectivas, atribuciones sobre los fines y roles asumidos por los distintos participantes en los procesos de evaluación.

Quimet, Bunnage, Carini, Kuh y Kennedy (2004) ejemplifican el uso de las entrevistas, los grupos de discusión y el juicio de expertos durante la validación de un cuestionario diseñado para conocer la satisfacción de los estudiantes con la docencia universitaria. Padilla, Gómez y Muñiz (2006) mostraron las distintas interpretaciones, valoraciones y atribuciones sobre los objetivos de las preguntas de un cuestionario de salud entre grupos de encuestados con y sin discapacidades. Presser, Rothger, Couper, Lessler, Martin, Martin y Singer (2004) realizan una exhaustiva revisión sobre la utilización de estos métodos para la elaboración y evaluación de cuestionarios incluidos sus aspectos consecuenciales.

Conclusiones

Debido a las dificultades conceptuales y metodológicas, la incorporación de la validación de las consecuencias del uso de los tests supone un reto considerable para los profesionales de la medición. Dificultades que atañen tanto a los contenidos como a la elección de los procedimientos adecuados para obtener evidencias creíbles con las que juzgar las consecuencias del uso de los tests. Tras realizar una presentación esquemática del esquema de validación basado en argumentos de validez (Cronbach, 1988; Kane, 1992, 2001, 2002), se ha mostrado su idoneidad para orientar la validación de las consecuencias a través de dos procesos de evaluación: la validación de las consecuencias de uso de los tests de «alto riesgo» en el contexto de la evaluación educativa y las repercusiones sobre la validez de las adaptaciones en los tests estandarizados para personas con discapacidades.

La construcción de un «argumento interpretativo» para iniciar el proceso de validación aporta dos claros beneficios en los estudios de validación: a) obliga a clarificar el contenido de la interpretación propuesta para las mediciones; y b) dirige la atención de la validación hacia los supuestos más débiles que soportan la interpretación deseada para el test. No se debe olvidar que el argumento interpretativo en su conjunto es tan débil como el más débil de los supuestos en que se basa (Kane, 2002). Respecto de la validación de las consecuencias, los beneficios de adoptar el esquema de validación basado en argumentos de validez son también evidentes: a) integra la validación de las consecuencias dentro de un único esquema de validación; y b) plantea la validación de las consecuencias en los mismos términos que la validación de las interpretaciones: formulación de la inferencia política, identificación de los supuestos en que se basa y búsqueda de evidencias para determinar si las consecuencias positivas del uso del test superan las potenciales consecuencias negativas. Sin ignorar los casos en los que el uso previsto para el test provoca la aparición de fuentes de varianza no relacionadas con el constructo o genera que el test no lo represente adecuadamente.

Tanto la evaluación de las consecuencias del uso de «tests de alto riesgo» en el contexto educativo, como la validación de las adaptaciones en los tests estandarizados para personas con discapacidades, sugieren consideraciones interesantes para analizar los procesos de validación del uso de los tests en el contexto español.

Aquí también se pueden identificar procesos de evaluación que implican la utilización de «tests de alto riesgo». Sobre todo si se acepta utilizar dicha etiqueta para aquellos tests que proporcionan mediciones con consecuencias directas importantes para los examinados, los programas o las instituciones implicadas en la evaluación (AERA, APA y NCME, 1999). Sin pretensión de exhaustividad, se podrían considerar como tales: las pruebas de acceso a la universidad, los exámenes para el acceso a la formación especializada en el campo de ciencias de la salud (e. g., exámenes MIR, PIR, etc.), el examen para la obtención del permiso de circulación, los procedimientos para la evaluación docente del profesorado universitario, etc. Todos estos ejemplos plantean el reto de realizar procesos de validación orientados tanto a la validación de las consecuencias como de las interpretaciones en que se basan. Sin contar con la necesidad de realizar estudios para obtener evidencias sobre potenciales consecuencias negativas de dichos procesos de evaluación: estrechamiento del currículo, focalización de la instrucción sólo en los contenidos evaluables por el test, estrategias de preparación para el test, etc. El esquema de validación basado en argumentos podría ser un enfoque para abordar la validación de estos procesos de evaluación.

La inclusión de un estándar sobre las cuestiones de justicia en el uso de los tests dentro de los *European Test User Standards for Test Use in Work and Organizational Setting* (EFPA y EAWOP, 2005), que incluye un apartado específico sobre las adaptaciones en los tests para personas con discapacidades, revela la creciente atención a estas cuestiones en nuestro contexto cultural más próximo. Sin duda puede considerarse una prueba más contundente la Orden Pre/1822/2006, de 9 de junio, por la que se establecen criterios generales para la adaptación de tiempos adicionales en los procesos selectivos para el acceso al empleo público de personas con discapacidad (BOE nº 140, 13/06/2006). Esta orden determina el tiempo extra que se debe añadir en las pruebas selectivas para personas con discapacidades, de forma proporcional al grado de discapacidad y para un amplio conjunto de tipos de deficiencias. Partiendo del avance que supone intentar eliminar las barreras que podrían impedir realizar mediciones válidas de los conocimientos y capacidades de las personas con discapacidades, no es difícil anticipar los interrogantes que plantean la extensión proporcional del tiempo de forma general sin contar con las características específicas de cada tipo de deficiencia y las demandas propias de las diferentes pruebas selectivas.

Sólo resta insistir en la necesidad de recurrir en los estudios de validación sobre las consecuencias del uso de los tests, procedimientos que permitan a las diferentes partes interesadas y agentes implicados en sus resultados participar en el proceso de validación. Procedimientos como las encuestas, los grupos de discusión o las entrevistas, entre otros posibles, permiten acordar las consecuencias relevantes sobre las que focalizar los estudios de validación y reconocer las evidencias necesarias para justificar el uso del test. De esta forma se podría lograr una mejor comunicación entre los profesionales de la medición y los agentes sociales relacionados con el uso de los tests.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Educación y Ciencia y los Fondos FEDER (Proyectos nº: SEJ2005-09144 y SEJ200-08924).

Referencias

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

Camilli, G. (2003). Comment on Cizek's «More unintended consequences of high-stakes testing». *Educational Measurement: Issues and Practice, 21,* 36-39.

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.

Cizek, G.J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20,* 19-27.

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

Cronbach, L.J. (1988). Five perspectives on validity argument. En H. Wainer y H.I. Braun (eds.): *Test validity.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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.

European Federation of Psychologist' Associations y European Association of Work and Organizational Psychology (2005). *European Test User Standards for Test Use in Work and Organizational Setting*. Disponible en la web: http://www.efpa.be/

Haladyna, T.M., y Downing, S.M. (2004). Construct irrelevant variances in high-stake testing. *Educational Measurement: Issues and Practice, 22,* 17-27.

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.

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

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.

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

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

Moss, P.A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22,* 13-25.

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

Nguyen, H.D., O'Neil, A., y Ryan, A.M. (2003). Relating test-taking attitudes and skill and stereotypes threat effects to racial gap in cognitive ability test performance. *Human Performance, 16,* 261-293.

Ouimet, J.A., Bunnage, J.C., Karini, R.M., Khu, G.D., y Kennedy, J. (2004).Using focus group, expert advice, cognitive interviews to establish the validity of a college student survey. *Research in Higher Education, 45,* 233-250.

Padilla, J.L., Gómez, J., Hidalgo, M.D., y Muñiz, J. (2006). La evaluación de las consecuencias del uso de los tests en la teoría de la validez. *Psicothema, 18,* 307-312.

Padilla, J.L., Gómez, J., y Muñiz, J. (2006). Assessment of overlap in construct by means of cognitive interview. Paper presented in 5th Conference of International Test Commission. Bruselas, Bélgica.

Pitoniak, M., y Royer, J. (2001). Testing accommodation for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71,* 53-104.

Presser, S., Rothger, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., y Singer, E. (eds.) (2004). *Methods for testing and evaluating survey questionnaires*. New Jersey: John Wiley & Sons.

Reckase, M.D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16.

Sireci, S.G. (2004). Validity issues in accommodation NAEP reading test (Center for Educational Assessment Research Rep. No. 515). Amherst, MA: School of Education, University of Massachusetts, Amherst.

Sireci, S.G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Research, 31,* 3-11.

Sireci, S.G. (2006). Test accommodations and test validity: Issues, research findings and unanswered questions. Paper presented in the Annual Meeting of National Center on Educational Outcomes Teleconference.

Sireci, S.G., Li, S., y Scarpati, S. (2003). The effects of test accommodations on test performance: A review of literature (Center for Educational Assessment Research Rep. No. 485). Amherst, MA: School of Education, University of Massachusetts, Amherst.