Psicothema 2011. Vol. 23, nº 4, pp. 701-706
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

PISA y el triángulo de la evaluación

María José Navas Ara[1] y Elsy Josefina Urdaneta Durán[2]
[1] Universidad Nacional de Educación a Distancia y [2] Universidad de Los Andes (Venezuela)

PISA es un conocido estudio internacional de evaluación educativa con 4 ediciones desde el año 2000. Este trabajo examina la validez de las medidas obtenidas en PISA utilizando como marco para el análisis el triángulo de evaluación. Para ello, se revisa la forma de proceder del estudio en los 3 vértices del triángulo: cognición, observación e interpretación. Este análisis revela la calidad de las medidas obtenidas y las bondades del diseño y procedimientos utilizados en el estudio, así como algún punto de mejora como, por ejemplo, el recurso a modelos de diagnóstico cognitivo en la fase de construcción de las pruebas y en el posterior análisis de datos.

PISA and the assessment triangle. PISA is a well-known and high profile Program for International Student Assessment with 4 editions since 2000. This study aims to examine the validity of PISA proficiency estimates, working with the framework provided by the assessment triangle. We pay explicit attention to how PISA proceeds as far as the three elements of the assessment triangle are concerned: cognition, observation, and interpretation. Results reveal not only the psychometrically sound proficiency estimates of PISA and the high standards reached, but also that there is room for improvement; for instance, cognitive diagnostic models could contribute both to test design and data analysis.

The Program for International Student Assessment, más conocido como PISA, es un estudio internacional de evaluación educativa promovido por la Organización para la Cooperación y el Desarrollo Económico (OCDE), en el que cada 3 años se evalúa a los alumnos que están a punto de finalizar la escolarización obligatoria en 3 áreas (lectura, matemáticas y ciencias). En cada edición del estudio se evalúa a fondo un área y se ofrece una pequeña actualización de las otras dos. En la cuarta edición (2009) la lectura ha vuelto a ser el área principal, con una participación de 65 países que representan en conjunto el 90% de la economía mundial.

En los estudios PISA se trabaja en base a lo que se considera necesario para el futuro ciudadano, principalmente para el acceso al mercado laboral: los conocimientos y destrezas evaluados no proceden prioritariamente del núcleo común de los currículos nacionales, sino de aquello que los expertos juzgan esencial para la vida adulta. PISA propone un concepto innovador y de difícil traducción al castellano (*literacy* / alfabetización) relacionado con la capacidad de los estudiantes de 15 años para aplicar conocimientos y destrezas adquiridas mayoritariamente en la escuela a situaciones similares a las que probablemente tendrán que enfrentarse después en su vida cotidiana, en particular, tiene que ver con su capacidad para analizar, razonar y comunicar de manera efectiva al plantear y resolver problemas en una variedad de áreas de conocimiento. La traducción que aquí se propone para el término *literacy* es competencia, dado el énfasis puesto en el uso funcional del conocimiento y habida cuenta de que la adquisición de estas capacidades es un proceso que dura toda la vida: PISA remite a la competencia o capacidad de los alumnos para enfrentarse a los retos de la vida adulta y para vivir en sociedad, proporcionando indicadores del capital en educación de las sociedades de los países participantes (Rico, 2006). De hecho, este estudio está destinado, en principio, a ser una herramienta útil para el diseño de políticas educativas que conduzcan a la calidad.

Este trabajo pretende examinar la validez de las medidas obtenidas en este estudio, utilizando un marco de análisis distinto al que se suele utilizar tradicionalmente en los estudios de validación.

Pellegrino, Chudowsky y Glaser (2001) sostienen que toda evaluación descansa en tres pilares: (1) un modelo sobre cómo representan los estudiantes el conocimiento en un área y cómo desarrollan su competencia en la misma; (2) las tareas o situaciones que permiten observar la forma de actuar de los estudiantes; y (3) un método para interpretar esas observaciones y poder realizar inferencias a partir de ellas. Estos pilares conforman los tres vértices de un triángulo (cognición, observación e interpretación), llamado triángulo de evaluación, que constituye un excelente marco de trabajo desde el que diseñar y evaluar instrumentos de medida. Como señalara certeramente Anastasi (1986), el proceso de validación de una prueba comienza con las primeras fases de la construcción del test, con la definición de su objetivo y la especificación del test: no es posible examinar las conexiones entre lo medido realmente y el constructo que se desea medir si éste no ha sido definido de forma apropiada con anterioridad y, a su vez, cuanto mejor y mayor haya sido la especificación del constructo de interés en las fases iniciales de construcción de la prueba, más abonado estará el camino para poder interpretar realmente las puntuaciones del test en términos de ese constructo, así como para utilizar dichas puntuaciones con el fin previsto. El triángulo de evaluación pro-

Fecha recepción: 19-7-10 • Fecha aceptación: 16-3-11
Correspondencia: María J. Navas
Facultad de Psicología
Universidad Nacional de Educación a Distancia
28040 Madrid (Spain)
e-mail: mjnavas@psi.uned.es

porciona un marco en el que se integran y engarzan de forma natural la construcción del test y la validación de sus puntuaciones: el modelo del vértice de la cognición proporciona una buena base para diseñar las tareas necesarias para obtener las observaciones de interés y, a su vez, ese modelo resulta también decisivo para el vértice de la interpretación.

En los siguientes apartados se procederá a examinar el trabajo realizado en el estudio PISA en cada uno de los vértices del triángulo para valorar en el apartado final la validez de las medidas obtenidas en este importante estudio.

Vértice de cognición

Una de las características que más diferencia el marco del triángulo de evaluación del modo habitual de operar al diseñar un estudio de evaluación es tener como punto de partida un modelo cognitivo. Este modelo describe cómo se representa el conocimiento y cómo se adquiere la maestría en un determinado campo y, además, señala cuáles son los aspectos más importantes del rendimiento de los estudiantes sobre los que obtener información, dando indicaciones sobre las tareas que podrían proporcionar esa información para realizar las correspondientes inferencias.

El proyecto DeSeCo (*Definition and Selection of Competencies*) constituye un antecedente indiscutible del marco de trabajo utilizado en PISA. Fue promovido por la OCDE en 1997 para determinar las competencias clave que necesitan las personas para poder adaptarse a un mundo caracterizado por el cambio, la complejidad y la interdependencia. Se identificaron 3 grandes categorías: la utilización interactiva de herramientas, la interacción en grupos heterogéneos y la actuación autónoma. Las competencias evaluadas en PISA pertenecen a esa primera categoría: la competencia lectora y matemática tendrían que ver con la utilización interactiva del lenguaje, textos y símbolos y la competencia científica con la utilización interactiva de información y conocimiento.

PISA define un marco de trabajo para conceptualizar cada una de estas competencias (véase OECD, 2009, para consultar los definidos en la cuarta edición del estudio). Dicho marco proporciona una definición que permite construir la prueba para obtener medidas de cada competencia en la población diana. Los marcos de trabajo para cada área son elaborados por sendas comisiones de expertos internacionales de los países de la OCDE participantes en el estudio y aprobados por los respectivos gobiernos.

Para describir estas competencias se utilizan tres dimensiones: (1) la situación en la que los estudiantes se enfrentan a problemas y aplican sus destrezas y conocimientos, que define ámbitos de problemas del mundo real; (2) el contenido necesario para resolver los problemas; y (3) los procesos que deben activarse para ello. Las dos primeras dimensiones tienen que ver con las características de los problemas a los que se enfrentan los estudiantes; la tercera dimensión apunta directamente a características de los sujetos. En la tabla 1 se recogen las definiciones propuestas para las tres competencias evaluadas y las categorías de estas dimensiones.

La definición y descripción de las competencias proporcionan el material necesario para elaborar las correspondientes matrices de especificación que, de algún modo, constituyen los planos para construir el edificio del test: indican cómo se va a representar empíricamente cada competencia, definiendo de manera precisa cuál ha de ser el contenido de la prueba y constituyen, por tanto, la base para el desarrollo de las tareas a incluir en los tests.

Vértice de observación

Se aborda aquí las tareas que se va a pedir a los estudiantes que realicen, así como los criterios para evaluar su actuación. Las tareas han de ser diseñadas de forma que eliciten tanto el conocimiento como los procesos que, según el modelo cognitivo del vértice anterior, son claves para la competencia del área en cuestión.

Para ilustrar la forma de proceder del estudio PISA en este vértice se trabajará con la competencia matemática (foco en la segunda edición del estudio).

Tabla 1
Caracterización de las tres competencias evaluadas en PISA*

Definición	Situación	Contenido	Proceso
COMPETENCIA LECTORA Capacidad para comprender, utilizar y analizar textos escritos para alcanzar los propios objetivos, desarrollar el conocimiento y potencial personales y participar en la sociedad	Privada Pública Laboral Educativa	*Textos continuos* *Textos discontinuos*	*Obtención de información* Comprensión global *Interpretación* *Reflexión sobre el contenido* *Reflexión sobre la forma*
COMPETENCIA MATEMÁTICA Capacidad para identificar y comprender el papel que juegan las matemáticas en el mundo, emitir juicios bien fundados y utilizarlas eficazmente para dar respuesta a las necesidades de una persona como ciudadano constructivo, comprometido y reflexivo	Personal Pública Científica Educativa/Laboral	*Cantidad* *Espacio y forma* *Cambio y relaciones* *Incertidumbre*	Reproducción Conexión Reflexión
COMPETENCIA CIENTÍFICA Capacidad para utilizar el conocimiento científico para identificar problemas, adquirir nuevo conocimiento, explicar fenómenos científicos, sacar conclusiones fundadas, comprender las características de la ciencia como forma de conocimiento e investigación y el papel que juegan en el entorno la ciencia y la tecnología y para implicarse como ciudadano reflexivo en temas relacionados con la ciencia	Personal Social Global	Conocimiento científico Conocimiento sobre la ciencia	*Identificación* *Explicación* *Utilización de pruebas*

* Se indica en cursiva las subescalas de las que se informa cuando se trata de la competencia foco en una edición del estudio

Confección de los ítems

Se fijó en 130 el número de ítems necesarios para evaluar la competencia matemática. Para poder contar con 130 ítems de la calidad psicométrica requerida y adecuadamente ajustados a la matriz de especificación de la prueba, convenía disponer inicialmente del triple de ítems, de los cuales se podrían pilotar unos 260 para llegar finalmente a 130.

La redacción de los ítems corrió a cargo de un consorcio internacional de agencias de investigación y desarrollo (ACER, Cito y NIER), que utilizaron distintos procedimientos de laboratorio cognitivo en el proceso; colaboraron también 15 países que hicieron sus propuestas de ítems.

Se redactaron preguntas de elección y construcción. Asimismo, se utilizó con frecuencia un enunciado común para diversas preguntas, lo que permite plantear tareas realistas que reflejan la complejidad propia de las situaciones en la vida real y hace posible también una utilización eficiente del tiempo de aplicación. Se pueden consultar ejemplos de preguntas en OCDE (2010).

Revisión inicial

Un panel de constructores profesionales de ítems examinaron en detalle todas las preguntas redactadas. Se utilizaron también procedimientos de laboratorio cognitivo en un buen número de ítems (técnicas de resolución en voz alta, entrevistas individuales y de grupo) para ver qué procesos utilizaban los estudiantes al realizar la tarea demandada.

Teniendo en cuenta esta información, el consorcio anterior de agencias de investigación y desarrollo seleccionó un conjunto de ítems que fueron prepilotados en Austria y en los países de las agencias del consorcio (Australia, Holanda y Japón) en escuelas con un número elevado de estudiantes de 15 años.

Analizados los datos de este pequeño estudio piloto, se puso en marcha el proceso de revisión en los países participantes. Se envió a cada coordinador nacional del estudio un formulario donde evaluar para cada ítem aspectos relacionados con su interés, relevancia para el currículo nacional y el marco PISA de trabajo, así como su adecuación cultural.

Los resultados obtenidos al analizar la información proporcionada por los coordinadores nacionales fueron utilizados en las reuniones que, de manera independiente, mantuvieron el grupo de expertos en Matemáticas y el denominado foro de Matemáticas, al que todos los países participantes podían enviar expertos.

El proceso de selección de los ítems que serían finalmente incluidos en el estudio piloto se inició con una reunión conjunta del foro y del grupo de expertos en Matemáticas, en la que los participantes valoraron la posible inclusión de cada ítem. En una reunión posterior con los coordinadores nacionales, el grupo de expertos propuso un conjunto de 237 ítems para el piloto.

Estudio piloto

Con toda la información recabada, el consorcio de agencias seleccionó un conjunto de 217 ítems, comenzándose entonces a trabajar en los países en la preparación de las versiones nacionales de los ítems seleccionados, llevando a cabo un proceso muy riguroso de traducción, adaptación y verificación externa para garantizar que todas las versiones fueran equivalentes. El proceso recomendado por PISA es la doble traducción desde dos versiones fuentes, una en inglés y otra en francés (OCDE, 2005).

Revisión final

Se llevó a cabo un detallado análisis de los datos obtenidos en el estudio piloto que incluyó también el análisis de las interacciones entre ítem y género e ítem y país, tras el cual se realizó una nueva ronda de valoraciones de los ítems en cada país. Los coordinadores nacionales prepararon un exhaustivo informe de revisión, identificando puntos fuertes y débiles de ítems concretos, relacionados con el proceso de traducción y verificación, con la preparación de las pruebas, la codificación de las respuestas o la grabación de los datos.

Seguidamente el grupo de expertos en Matemáticas revisó toda la información disponible y preseleccionó 88 ítems, sobre los que el consorcio todavía hizo algunos ajustes. La selección final de ítems fue presentada para su aprobación al consejo de gobierno de PISA y a una reunión con los coordinadores nacionales.

Confección de la prueba

El balance final de ítems después de todo el proceso fue 85 de Matemáticas, 28 de competencia lectora, 35 de competencia científica y 19 de solución de problemas (área incluida únicamente en 2003).

Con ellos se formaron 7 bloques de ítems de matemáticas, 2 de lectura, 2 de ciencias y 2 de solución de problemas (con 30 minutos para responder a cada bloque). Se determinó que cada estudiante debía responder a 4 bloques (2 horas de aplicación), utilizando para ello un conocido diseño de muestras matriciales múltiples (Bloques Incompletos Balanceados). Según este diseño, los ítems se combinan en bloques y los bloques en cuadernillos, que son asignados aleatoriamente a los estudiantes. Los cuadernillos se diseñan de modo que cada bloque de ítems debe aparecer una vez en cada una de las posibles posiciones dentro del cuadernillo, lo que da un total de 13 cuadernillos distintos (tabla 2).

Tabla 2
Configuración de los cuadernillos administrados en el estudio PISA 2003

Cuadernillo	Bloque 1	Bloque 2	Bloque 3	Bloque 4
1	M1	M2	M4	L1
2	M2	M3	M5	L2
3	M3	M4	M6	SP1
4	M4	M5	M7	SP2
5	M5	M6	C1	M1
6	M6	M7	C2	M2
7	M7	C1	L1	M3
8	C1	C2	L2	M4
9	C2	L1	SP1	M5
10	L1	L2	SP2	M6
11	L2	SP1	M1	M7
12	SP1	SP2	M2	C1
13	SP2	M1	M3	C2

C= Ciencias; L= Lectura; M= Matemáticas; SP= Solución de problemas

Tras la fase de recogida de datos con las pruebas así construidas, se procedió a la codificación y corrección de las respuestas de los estudiantes. Para las preguntas de construcción se trabajó con 4 correctores nacionales y algunas fueron también examinadas por un quinto (en algunos casos incluso por un sexto) corrector independiente, para estimar el posible sesgo en la corrección en los distintos países.

Vértice de interpretación

Éste es el vértice con una relación más directa con el análisis tradicional de la validez de las puntuaciones, medidas u observaciones obtenidas al pedir a los sujetos que realicen las tareas diseñadas según el modelo que desde el vértice de la cognición define y explica la competencia en cuestión. Según se recoge en la última edición de los *Standards for Educational and Psychological Testing*, la validez se refiere justamente al grado en que tanto los datos como la teoría apoyan la interpretación de las puntuaciones de un test vinculada al uso propuesto para éste (AERA, APA y NCME, 1999).

El trabajo realizado en PISA para interpretar las estimaciones de la competencia de los estudiantes ha sido ingente. Así, llevó más de un año todo el trabajo preparatorio para informar de los resultados en la primera edición del estudio, participando junto al consorcio de agencias los grupos de expertos de cada área, el consejo de países participantes, los coordinadores nacionales y el grupo PISA de asesoría técnica. El proceso seguido en todas las ediciones del estudio es el siguiente.

Una vez codificadas y corregidas las respuestas de los estudiantes, para estimar la competencia de cada alumno se utiliza el modelo logit multinomial de coeficientes aleatorios (Adams, Wilson y Wang, 1997) y se trabaja con la metodología de valores plausibles, dado el tipo de diseño utilizado para recoger los datos.

En todas las ediciones del estudio se ofrecen estimaciones de la competencia en las 3 áreas evaluadas; para el área foco se estima la competencia no solo en una escala general, sino también en algunas subescalas, definidas según las dimensiones establecidas en el marco PISA de trabajo y subrayadas en la tabla 1. Para validar las estimaciones en estas subescalas se utiliza una doble estrategia: (1) el recurso al juicio de expertos, para ver si los ítems son asignados a las subescalas definidas; y (2) un amplio proceso de escrutinio por parte de los coordinadores nacionales para evaluar hasta qué punto son escalas verdaderamente informativas.

Para facilitar la interpretación de las estimaciones obtenidas para los estudiantes se transforma dichos valores a una nueva escala con media 500 y desviación típica 100, que corresponden a la media y desviación típica de los países OCDE del estudio. Esto se hace en la primera edición del estudio, en que una competencia es foco para establecer la escala que servirá de base o referencia para examinar dicha competencia a lo largo de las distintas ediciones del estudio, utilizando para la equiparación un diseño de anclaje de ítems. Por consiguiente, la escala base para las 3 competencias es idéntica (500, 100) pero corresponde a distintas ediciones del estudio. Así, la competencia lectora mostrada por los estudiantes españoles en la edición de 2009 (481 puntos) hay que referirla a la escala establecida en 2000, su competencia matemática (483 puntos) a la escala establecida en 2003 y su competencia científica (488 puntos) a la escala establecida en 2006 (OCDE, 2010). Además, el carácter cíclico del estudio permite realizar un seguimiento regular de la evolución de las competencias básicas de los estudiantes.

Con el fin de mejorar la comunicación de los resultados y conseguir que la información proporcionada resulte verdaderamente útil y accesible a educadores y responsables políticos, así como a la opinión pública, PISA define distintos niveles de competencia convirtiendo las anteriores estimaciones de la competencia de los estudiantes en variables categóricas, con 8 categorías para la competencia lectora y 6 para la matemática y científica en la edición 2009. De este modo, se informa no solo en términos cuantitativos, sino que se cualifica esa información: cada nivel de competencia va acompañado de su correspondiente descripción de destrezas asociadas, con una indicación concreta de qué es lo que se espera de un estudiante cuya competencia se localiza justamente en ese nivel. La teoría de respuesta al ítem presta aquí un servicio impagable, al estar en la misma escala de medida las estimaciones de la competencia de los estudiantes y las estimaciones de la dificultad de los ítems de la prueba utilizada para evaluar esa competencia. Por tanto, la diferencia entre el valor estimado para un sujeto en la característica evaluada y la dificultad del ítem tiene un significado directo en términos de la probabilidad de responder correctamente a cada ítem.

La definición de los niveles de competencia se basa justamente en esta propiedad, así como en el análisis de las demandas cognitivas de los ítems y en la observación de cómo responden los estudiantes a estos ítems. En particular, para cada área evaluada los niveles de competencia se definen de forma que: (1) los estudiantes con una competencia estimada en el límite inferior de un determinado nivel han de ser capaces de responder correctamente al menos a la mitad de los ítems cuya dificultad estimada se encuentra justamente en ese nivel; y (2) los estudiantes con una competencia estimada en el límite superior de un determinado nivel han de ser capaces de responder correctamente en torno al 80% de los ítems cuya dificultad estimada se encuentra en ese nivel. Lo anterior debe ser válido para todos los niveles, lo que implica que los distintos niveles tienen que tener aproximadamente la misma amplitud: 0.8 logits o, lo que es lo mismo, una probabilidad de 0.62 de responder correctamente a los ítems de ese nivel.

El nivel de competencia al que se asigna a un determinado estudiante es al nivel más alto en el que se espera que responda correctamente a la mayoría de los ítems. Más específicamente, al nivel en cuyo intervalo se encuentra la puntuación en la escala que corresponde para ese estudiante a una probabilidad de 0.62 de responder correctamente a ítems cuya dificultad estimada se localiza en dicho intervalo.

Para describir cada uno de los niveles definidos y poder informar acerca de lo que puede hacer un estudiante con una competencia estimada en ese nivel es preciso examinar las características de los ítems cuya dificultad coincide con el intervalo de la escala definido para ese nivel, ya que esto proporcionará el perfil típico de destrezas y habilidades de que dispone un estudiante en ese nivel de la escala. El lector interesado en la descripción de los niveles de competencia de la escala general y de las subescalas definidas para las áreas evaluadas puede acudir a los informes correspondientes de los estudios PISA: OECD (2005) para la competencia matemática, OECD (2007) para la científica y OECD (2010) para la lectora.

Discusión

El resultado de la radiografía del estudio PISA, proporcionada por el triángulo de evaluación, evidencia una buena salud de las medidas obtenidas con las pruebas utilizadas en esta serie de estudios. Pese a no partir de un modelo de cognición y aprendiza-

je explícitamente formulado, el marco de trabajo utilizado indica claramente qué aspectos evaluar en cada competencia y proporciona indicios suficientes para diseñar las tareas que servirán para obtener observaciones que harán posible realizar las inferencias pertinentes acerca del grado de competencia de los alumnos en los países participantes. Por lo que respecta al vértice de observación, el modo de proceder en el estudio PISA es un ejemplo de buenas prácticas a la hora de construir ítems que pueden proporcionar buenas observaciones o medidas: se plantean tareas que emanan directamente de la matriz de especificación de la prueba, se somete a comprobación empírica los procesos que estas tareas ponen en juego (mediante el análisis cognitivo de éstas), se utilizan variados formatos de preguntas que permiten evaluar los distintos contenidos, situaciones y procesos recogidos en la matriz de especificación y, gracias al diseño de muestras matriciales múltiples utilizado, es muy importante el grado de cobertura conseguido para el área principal evaluada en cada edición del estudio. Por último, es verdaderamente destacable el esfuerzo realizado en el vértice de interpretación, para informar de los resultados en términos de escalas bien ancladas teóricamente y fácilmente interpretables desde el punto de vista de la política educativa y también para todas las audiencias interesadas, con una referencia clara a lo que los estudiantes son capaces de hacer.

El triángulo de evaluación es una propuesta realizada desde el *National Research Council* por el comité encargado de revisar los avances acaecidos en el campo de las ciencias cognitivas y de la medición para examinar sus implicaciones de cara a la evaluación educativa. El informe final concluía con 12 recomendaciones para ampliar la integración entre el campo de la cognición y la medición. El estudio PISA trabaja en la línea de al menos 3 de las recomendaciones sugeridas por el comité, al facilitar la comprensión de los principios básicos en la utilización e interpretación de las medidas obtenidas con los tests (recomendación 12), al promover nuevas maneras de informar sobre los resultados, que transmitan diferencias importantes en los distintos niveles de competencia y que resulten claras para las audiencias implicadas (recomendación 8) y al hacer posible la medición del cambio o progreso en el tiempo de estudiantes y sistema educativo y permitir realizar un análisis multinivel de los factores relacionados con dicho cambio (recomendación 10).

En el número extraordinario que la Revista de Educación dedicó al estudio PISA, Martínez Arias (2006) sigue las directrices marcadas por el *Board on International Comparative Studies in Education* (National Research Council, 1990) para valorar la calidad metodológica de la edición 2003, concluyendo que «sigue todas las buenas prácticas de las evaluaciones internacionales en cuanto a representatividad de las muestras, calibración adecuada de los parámetros, estimaciones precisas de los resultados, buenas traducciones, adecuados cuestionarios de contexto bien codificados y un plan adecuado de informes dirigidos a diversas audiencias» (p. 124).

En suma, la calidad de las medidas obtenidas con las pruebas utilizadas en PISA resiste bastante bien distintos marcos de análisis, aunque siempre hay espacio para la mejora. Así, sería deseable contar con un modelo de cognición y aprendizaje para apuntalar en el estudio el vértice de la cognición: una forma excelente de mejorar la calidad de los ítems de un test, así como la validez de las inferencias realizadas a partir de sus puntuaciones, consiste justamente en diseñar los ítems con arreglo a un modelo cognitivo, que explique con claridad qué conocimientos y destrezas utilizan habitualmente los estudiantes en un área determinada para resolver problemas o tareas estandarizadas (Borsboom, 2005; Cui y Leighton, 2009; Embretson y Gorin, 2001; Leighton, 2004; Mislevy, 2006; Nichols, 1994). Esto tendría un impacto evidente en el vértice de la observación y, además, un efecto colateral muy positivo en el vértice de la interpretación, ya que serviría para corregir una limitación del estudio PISA relacionada con su escasa utilidad para reorientar el trabajo del profesor en el aula, al ser un estudio básicamente dirigido a la comparación entre países con una marcada orientación política. Dado que cada día es mayor el esfuerzo que se pide a los centros escolares para que participen en evaluaciones educativas (regionales, nacionales e internacionales), podría resultar motivador y, sobre todo, muy útil proporcionarles información específica acerca de puntos fuertes y débiles de los alumnos, esto es, información relevante para mantener o corregir el rumbo dentro del centro, más allá de la imagen que ofrece el estudio sobre la situación de una región o país en relación a la zona OCDE y/o sus países de referencia. El recurso a modelos de diagnóstico cognitivo (Gorin, 2006; Leighton y Gierl, 2007; Rupp y Templin, 2008; Rupp, Templin y Henson, 2010) en la fase de construcción de las pruebas y en el posterior análisis de datos puede proporcionar esa información diagnóstica de utilidad para la práctica en el aula que relacione los resultados del estudio con el trabajo del profesor.

Referencias

Adams, R.J., Wilson, M., y Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

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

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

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

Cui, Y., y Leighton, J.P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 429-449.

Embretson, S., y Gorin, J.S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343-368.

Gorin, J.S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21-35.

Leighton, J.P. (2004). Avoiding misconceptions, misuse and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(2), 6-15.

Leighton, J.P., y Gierl, M.J. (2007). Why cognitive diagnostic assessment? En J.P. Leighton y M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.

Martínez Arias, M.R. (2006). La metodología de los estudios PISA. *Revista de Educación, nº extraordinario*, 111-129.

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. En R. Brennan (Ed.), *Educational Measurement*. Westport, CT: Praeger and American Council on Education.

Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessment. *Review of Educational Research, 64*, 575-603.

National Research Council (1990). A framework and principles for international comparative studies in education. Washington, DC: National Academy Press.

OECD (2005). *PISA 2003 Technical Report*. París: OCDE.

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World*, vol. 1. París: OECD.

OECD (2009). *PISA 2009 Assessment Framework. Key competencies in reading, mathematics and science*. París: OECD.

OECD (2010). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science* (Volume I). Recuperado el 8 de diciembre de 2010 de http://dx.doi.org/10.1787/9789264091450-en.

Pellegrino, J.W, Chudowsky, N., y Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Rico, L. (2006). Marco teórico de evaluación en PISA sobre matemáticas y resolución de problemas. *Revista de Educación, extraordinario 2006*, 275-294.

Rupp, A.A., y Templin, J.L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state of the art. *Measurement, 6*, 219-262.

Rupp, A.A., Templin, J.L., y Henson, R.J. (2010). *Diagnostic measurement: Theory, methods and applications*. Nueva York: Guilford Press.