La metodología de los estudios PISA

Rosario Martínez Arias
Universidad Complutense

Resumen:

La OCDE ha llevado a cabo un gran esfuerzo en una serie de estudios conocidos como *The Programme for International Student Assessment* (PISA). Probablemente, PISA-2003 es uno de los más amplios y complejos estudios internacionales de rendimiento educativo. Este artículo describe los aspectos metodológicos de PISA-2003 y resume las principales actividades en el desarrollo de los instrumentos de recogida de datos, traducción de los instrumentos de evaluación, diseño muestral, análisis de datos y la presentación de los resultados. Se evalúa la calidad metodológica del estudio desde los criterios establecidos por el *National Research Council*. Se concluye destacando los esfuerzos realizados para obtener buenas tasas de respuesta y en el diseño de los instrumentos, así como la atención prestada a las lecciones aprendidas de estudios anteriores. Aunque PISA es metodológicamente correcto en todos los frentes, quedan algunas cuestiones no resueltas, como la falta de motivación de los alumnos en las evaluaciones sin consecuencias, el rigor en el control de las tasas de respuesta y exclusiones, la equidad y neutralidad en la investigación y el uso e impacto de los resultados.

Palabras clave: evaluación internacional, diseño de tests, teoría de la respuesta al ítem, valores plausibles, pesos de muestreo, niveles de rendimiento, informes.

Abstract: *The methodology of PISA studies*

The OECD has invested heavily in a series of studies known as «The Programme for International Student Assessment (PISA)». Probably, PISA-2003 is one of the most largest and complex international educational surveys. This paper describes the technical aspects of PISA-2003 and summarizes the main activities involved in the development of the data collection instruments, test translation, sample design, data analysis, and reporting of the results. We assess the methodological quality from the point of view of National Research Council's criteria. Considerable efforts have been made to obtain good response rates and careful attention has been devoted to the design of instruments, and many lessons from previous studies were clearly absorbed. PISA is technically sound on virtually all fronts, but there are some questions unanswered such as the student's motivation in low-stake assessments, the rigor in fulfilling of rates of response and exclusions, research neutrality and fairness and the use and impact of the results.

Key words: international assessment, test design, Item Response Theory, plausible values, sampling weights, achievement levels, reporting.

LAS EVALUACIONES INTERNACIONALES DEL RENDIMIENTO EDUCATIVO Y SUS DIFICULTADES METODOLÓGICAS

El estudio PISA representa la primera incursión de la Organización para la Cooperación y Desarrollo Económico (OCDE) en el complejo mundo de las evaluaciones internacionales, prácticamente lideradas desde sus comienzos en 1964 por la *International Association for the Evaluation of Educational Achievement* (IEA). Puede considerarse que la idea latente en todos estos estudios es la de considerar el mundo como una especie de *laboratorio educativo global*, donde diferentes políticas y prácticas educativas nacionales producen distintos resultados. Se considera que las diferencias en el rendimiento de los alumnos pueden ligarse a características de los sistemas educativos, aunque éstas deben interpretarse con cautela, explorando los correspondientes contextos económicos y culturales. Los propósitos de estos estudios son variados, pero parece haber bastante acuerdo en una serie de puntos (Beaton, Postlethwaite, Ross, Spearritt y Wolf, 1999; National Research Council, 1990; Mislevy, 1995; Postlethwaite, 1999). Un propósito esencial es el uso de los resultados como «criterio de referencia» (*benchmarking)*, ya que el examen de las políticas y prácticas educativas de los países de altos rendimientos puede proporcionar nuevas pautas para la mejora a los países participantes. Para ello, se exploran asociaciones entre diferentes políticas y prácticas y el rendimiento, así como con otras variables contextuales.

El creciente interés prestado por los distintos países a estos estudios (su número se ha multiplicado en los últimos años) procede de la convicción generalizada de que la productividad económica futura de un país depende de los altos niveles de conocimientos y competencias en la población activa y de que mostrarán una mejor posición internacional aquellas naciones que tengan población con mayor nivel, especialmente en lectura, matemáticas y ciencias. Hay algunas evidencias de relación positiva entre los niveles de rendimiento educativo y desarrollo de la productividad nacional (OCDE, 1996; Beaton et al., 1999), pero existe un fuerte debate sobre la dirección de la causalidad (Levin, 1988; Robinson, 1999). También en el nivel individual se han encontrado relaciones entre resultados educativos, calidad del empleo e ingresos (OCDE, 1989). Un dato que apoya esta fuerte creencia nos lo proporciona la propuesta de indicadores educativos de la UE para Europa 2010, en la que varios de los indicadores proceden del estudio PISA en sus sucesivas oleadas (European Commission, 2002, 2004). En este sentido, muchos observadores creen que las comparaciones internacionales están llevando al desarrollo de objetivos y estándares más ambiciosos para el aprendizaje de los estudiantes.

Los estudios comparativos internacionales son complicados y difíciles de llevar a cabo, ya que suelen tener varias finalidades y están destinados a informar a audiencias múltiples. En general, su calidad metodológica es alta y por tanto sus resultados pueden tomarse con bastante credibilidad. Cuatro décadas de experiencia con este tipo de evaluaciones, así como los desarrollos generados en torno al NAEP (*National Assessment of Educational Progress*, evaluación de EEUU,

iniciada en 1969 y llevada a cabo con estricta periodicidad), han llevado a sustanciales mejoras metodológicas en todos los aspectos que van desde la construcción de los tests hasta la presentación de los resultados (Porter y Gamonal, 2002). Los avances metodológicos aparecen muy bien recogidos por la IEA en sus *Technical Standards* (Gregory y Martin, 2001; Martin, Rust y Adams, 1999). Además, los estudios suelen estar coordinados por consorcios formados por las entidades más prestigiosas en estas evaluaciones. La participación de un país le permite familiarizarse con estas metodologías, lo que le resultará beneficioso para llevar a cabo sus propias evaluaciones del sistema educativo.

A continuación, se examinan los principales aspectos de la metodología de PISA-2003, aunque por razones de espacio estamos obligados a una revisión muy breve que simplemente permita a los lectores hacerse una idea de la calidad metodológica del estudio. Las personas interesadas en una descripción completa pueden examinar los informes técnicos publicados en 2002 y 2005 (Adams y Wu, 2002; OCDE, 2005a) y los *Data Analysis Manuals* (OCDE, 2005b; OCDE, 2005c). Todos ellos se pueden obtener en formato PDF de http://www.oecd.org.

LOS INSTRUMENTOS DE EVALUACIÓN UTILIZADOS EN PISA

El desarrollo de los marcos

Todo instrumento de medida debe comenzar con la identificación de su propósito y de las inferencias que se pretenden hacer a partir de sus puntuaciones (AERA, APA y NCME, 1999; Millman y Greene, 1989). Esta tarea es especialmente complicada en los estudios internacionales, ya que debe mantenerse el principio de equidad que debe presidir toda evaluación, por lo que en la definición de los marcos deben tenerse presentes los currículos nacionales y la «oportunidad para aprender» (Floden, 2002) que han tenido las diferentes poblaciones implicadas. Linn (2002) habla de un continuo cuyos extremos son la «unión» y la «intersección» de currículos, encontrándose habitualmente los marcos más próximos a la intersección. Algunos estudios, no obstante, ponen de relieve escasas variaciones en la posición de los países cuando los alumnos responden a tareas más próximas a su currículo (Greaney y Kellagaham, 1996). En PISA, esta cuestión ha resultado menos problemática, ya que el objetivo va más allá del currículo, poniendo el acento en destrezas o competencias más generales que deberían poseer todos los jóvenes a los 15 años, cuando finalizan o están a punto de finalizar la escolaridad obligatoria, para poder hacer frente a los retos de la sociedad. El dominio de contenido se define por la opinión de los expertos internacionales y no por el currículo.

El diseño del test comenzó con el desarrollo de los marcos para cada uno de los dominios cognitivos evaluados, que fueron publicados por los responsables del estudio (OCDE, 2003). Los diferentes aspectos del marco se recogen en las denominadas «matrices de especificaciones», que representan la base para el desarrollo de las tareas que se deben incluir en los tests y que son producto del trabajo de los diversos países.

La escritura de ítems y los estudios piloto

Todos los países fueron invitados a participar en esta tarea, recibiéndose ítems de 15 países, que fueron revisados y adaptados por los responsables del desarrollo de ítems (ACER, CITO y Universidad de Leeds). Esta colaboración internacional representa un esfuerzo por hacer más neutral y equitativa la evaluación.

En un intento de mejorar la validez de contenido y de constructo de los ítems, el estudio PISA utilizó las siguientes metodologías en las primeras fases de análisis:

- Paneles de expertos para la revisión de los ítems, en una primera fase de los responsables del desarrollo de los instrumentos y posteriormente paneles de expertos internacionales.
- Entrevistas cognitivas con alumnos individuales y grupos de alumnos.
- Tests pilotos con muestras de estudiantes representativas de la población objetivo de Australia, Japón, Holanda y Austria. A veces, se presentaron versiones alternativas del mismo ítem para identificar la mejor forma de presentación.

Cada ítem fue sometido a un estudio piloto inicial en escuelas con un número elevado de estudiantes de la edad de la población objetivo (en Australia, Japón, Holanda y Austria). Algunas tareas fueron presentadas en versiones alternativas para determinar la mejor forma de presentación. Los datos fueron analizados con técnicas clásicas de análisis de ítems.

La principal novedad de PISA en esta fase fue el uso en el análisis de ítems de los procedimientos de la denominada «psicometría cognitiva». Las recomendaciones más actuales insisten en este aspecto como una de las claves para lograr la validez de constructo de las tareas, ya que permiten disponer de una especie de «modelo del alumno en la resolución de la tarea» que es la mejor forma de comprender sus demandas cognitivas (Mislevy, Steinberg y Almond, 2003). Se utilizaron técnicas de resolución en voz alta (*thinking aloud*), entrevistas individuales y entrevistas de grupo.

Otro aspecto destacable de los ítems o tareas de PISA es la inclusión de un importante número de ítems de respuesta abierta y extendida, tendencia mostrada también en otras evaluaciones internacionales recientes como el TIMSS (Martin, Gregory y Stemler, 2000; Martin, Mullis, y Chrostowski, 2004). La inclusión de este tipo de ítems puede reducir la eficiencia del test en cuanto a facilidad de corrección, fiabilidad de las puntuaciones y tiempo requerido para la respuesta (Linn, 2002); no obstante, parece imprescindible para evaluar procesos superiores. En las páginas 28 a 30 del *Technical Report* (OCDE, 2005a), puede verse la distribución de los ítems por formato, contenidos y demandas cognitivas.

Dado el elevado número de tareas que requieren codificación subjetiva por parte de evaluadores formados, también se llevó a cabo la formación de codificadores. Una aproximación novedosa fue el uso de técnicas cualitativas del análisis de los datos, como el «escalamiento dual» (Nishisato, 1980), también conocido como «análisis de correspondencias» (Gifi, 1990), como ayuda en esta tarea.

Además, se llevaron a cabo los cálculos habituales de fiabilidad entre jueces o codificadores, con objeto de eliminar los ítems más problemáticos.

Finalmente, tal como se recomienda en los *IEA Technical Standards* (Martin et al., 1999) se analizaron las principales propiedades psicométricas de los ítems, con los datos derivados de muestras de alumnos. Los cálculos fueron realizados con el programa *ConQuest* (Wu, Adams y Wilson, 1997).

La distribución de los ítems seleccionados en cuanto a contenidos, demandas cognitivas y formatos puede consultarse en la página 24 del *Technical Report* (OCDE, 2005a).

La traducción de las pruebas cognitivas

La necesidad de traducir-adaptar los tests a múltiples lenguas es una importante fuente de problemas en las evaluaciones internacionales. Las traducciones pueden producir ítems que difieran en dificultad y demandas cognitivas entre los diferentes idiomas y el rendimiento de los alumnos puede variar mucho por cambios aparentemente menores en la traducción, lo que llevaría a comparaciones no válidas (Hambleton, 2002; Hambleton, Merenda y Spielberger, 2005).

En los procesos de traducción-adaptación de instrumentos en general y en las evaluaciones internacionales en particular, son habituales los procedimientos de «back translation» (nueva traducción a la lengua original) y la doble traducción con puesta en común por una tercera persona (procedimiento usado en TIMSS). Una importante novedad de PISA muy bien valorada por los expertos (Hambleton, 2002) fue la introducción de dos versiones fuente, en inglés y francés, lo que lleva a la intervención y al acuerdo de cuatro personas. En el *Technical Report* del estudio del 2000 (Adams y Wu, 2002), se exploraron las consecuencias del uso de las dos lenguas fuente, encontrando que era una buena aproximación. La tarea de la traducción se completó con otras tareas: solicitar a los países la inclusión de notas de todos los posibles problemas de adaptación, desarrollo de guías detalladas de traducción y adaptación del material del test y de la revisión después del estudio piloto, formación de los equipos nacionales, y selección y formación de un grupo de verificadores internacionales.

Estudio piloto en los países participantes

Para el estudio piloto se elaboraron cuadernillos en todos los países participantes con el mismo diseño que se seguiría en la evaluación efectiva. Con los datos de estos pilotos se analizaron las propiedades psicométricas de los ítems (índices de dificultad, discriminación, opciones de respuesta, análisis de distractores, tasas de omisiones en los ítems, funcionamiento diferencial entre países...) utilizando el programa *ConQuest* (Wu et al., 1997). También se analizó el tiempo requerido para responder al test y se llevó a cabo un estudio de fiabilidad de los codificadores. Los resultados se enviaron para su examen a los coordinadores nacionales, que además debían valorar los ítems en función de los objetivos del estudio.

En la selección final de ítems, realizada por expertos, se tuvieron en cuenta las propiedades psicométricas, la valoración de los ítems por los grupos nacionales, información sobre la traducción y las restricciones impuestas por el marco en cada dominio. La composición definitiva del conjunto de ítems puede verse en las páginas 28-30 del *PISA-2003 Technical Report* (OCDE, 2005b).

Estudios de fiabilidad de los codificadores

Para el estudio de fiabilidad se seleccionaron 16 codificadores por país, siendo asignados los cuadernillos de forma aleatoria a los codificadores según diseño que se puede ver en la p. 92 del *PISA-2003 Technical Report* (OCDE, 2005a). También se llevó a cabo una labor de codificación por múltiples calificadores para asegurar mejor la fiabilidad a partir de una muestra de 900 cuadernillos.

Los análisis de fiabilidad de los codificadores se realizaron en el marco del modelo psicométrico de los componentes de la varianza o Teoría de la Generalizabilidad (Brennan, 2002; Cronbach, Gleser Nanda y Rajaratnam, 1972; Martínez Arias, 1995; OCDE, 2005a), que permite estimar las proporciones de varianza de las puntuaciones debidas a diferentes efectos, entre otros los codificadores. Se encontró que todos los componentes de la varianza en los que estaban implicados los codificadores (codificadores y todas las interacciones en las que intervienen) son muy pequeños, concluyendo que no hay efectos sistemáticos debidos a los calificadores. Se obtuvieron los coeficientes de generalizabilidad (una forma de coeficiente de fiabilidad) para cada uno de los países. En el capítulo 14 del *Technical Report* (OCDE, 2005a), se presentan los resultados por países.

La construcción de los tests o cuadernillos para la evaluación

Para que los tests de evaluación del rendimiento tengan validez de contenido y se ajusten bien a las matrices de especificaciones, es preciso incluir muchos ítems. Por razones prácticas, es imposible y seguramente poco deseable, examinar a todos los alumnos con todos los ítems: 1) después de mucho tiempo cumplimentando el test los sujetos empiezan a verse afectados por la fatiga y 2) habría menor tasa de participación de los centros, sesgando los resultados y aumentado el error muestral. Para eliminar estos problemas, manteniendo las necesidades de cobertura del dominio evaluado, los alumnos son asignados aleatoriamente a subconjuntos del conjunto de ítems; es decir, solamente algunas submuestras de estudiantes responden a cada uno de los ítems. Este diseño se conoce en la literatura psicométrica como «muestreo matricial de ítems» (*Matrix Sampling of Items*) (Beaton, 1997; Childs y Jaciw, 2003) y es el que se utiliza en la mayor parte de las evaluaciones a gran escala. La forma más común de construcción de cuadernillos o formas de test es el diseño BIB (*Balanced Incomplete Blocks*) en el que los ítems son divididos en varios bloques, que se combinan en cuadernillos y cada bloque es emparejado una vez al menos con todos los demás bloques. En PISA-2003, el conjunto de ítems fue dividido en 13 bloques y la asignación de ítems a los bloques tuvo en cuenta la dificultad esperada de los ítems

y el tiempo de respuesta necesario. Se construyeron 13 cuadernillos, cada uno de ellos con cuatro bloques. Una innovación de PISA-2003 es que cada bloque aparece una vez en una de cuatro posiciones, para controlar algunos efectos de la posición de los bloques que se encontraron en PISA-2000. Una disposición de la estructura de cuadernillos y bloques puede verse en la página 68 del *Technical Report* (OCDE, 2005a). El diseño tiene el inconveniente de que dificulta la comparabilidad de las puntuaciones, porque los sujetos responden a diferentes cuadernillos y éstos no son estrictamente paralelos. Esta dificultad técnica puede solventarse utilizando un modelo de medida de los denominados de «Teoría de Respuesta al Ítem (TRI)», muy bien documentados en la literatura psicométrica (Hambleton, Swaminathan y Rogers, 1991; Lord, 1980; Martínez Arias, 1995; Thissen y Wainer, 2001; Van der Linden y Hambleton, 1997). Los bloques comunes a diferentes cuadernillos o de enlace permiten la calibración conjunta de los ítems y la comparabilidad de las puntuaciones. Este tipo de equiparación basada en ítems o bloques comunes es la base también de los estudios de tendencias, en los que se usan ítems de enlace o anclaje de evaluaciones anteriores. La equiparación de puntuaciones procedentes de diversas formas de tests, sea en el nivel individual o en niveles agregados, ha sido muy tratada en la psicometría (Fitpatrick, Lee y Gao, 2001; Kolen y Brennan, 2004; Linn, 1993; Mislevy, 1992).

El uso de modelos de Respuesta al Ítem es útil no solamente para resolver los problemas de comparabilidad, sino también porque permiten crear una escala continua de «habilidad» o competencia, en la que se pueden situar simultáneamente sujetos e ítems, lo que facilita los informes de resultados. Existen muchos modelos de este tipo en la literatura, pero básicamente pueden clasificarse en dos grandes bloques: los que derivan del modelo de Rasch (1960/1980), que tienen propiedades óptimas de escalamiento, y los que derivan de la línea de Lord (1980), que buscan una mayor adaptación a las respuestas empíricas de los sujetos e incluyen más parámetros para explicar las respuestas a los ítems. Parece que la decisión de utilizar unos u otros modelos va más allá de aspectos puramente técnicos, siendo las posturas de sus partidarios bastante irreconciliables (Thissen y Orlando, 2001). En los modelos de Rasch, el único factor que influye en la probabilidad de acierto es la distancia entre la dificultad del ítem (o de la categoría de respuesta) y la habilidad del sujeto. También, la dificultad relativa de un ítem resulta de la comparación del ítem con los otros ítems y es independiente de la habilidad de los estudiantes.

El modelo concreto aplicado en PISA es una forma generalizada del modelo de Rasch, el modelo de *coeficientes mixtos* (Adams, Wu y Wilson, 1997), que admite la multidimensionalidad (necesaria para los diferentes dominios y subdominios de PISA) y en el que los ítems son descritos por medio de un conjunto de parámetros desconocidos y la habilidad de los alumnos es un resultado aleatorio. El modelo permite el análisis de interacciones del ítem con otras variables, así como la posibilidad de generar para cada sujeto una distribución posterior con múltiples resultados. También permite especificar modelos diferentes dentro del mismo test para distintos tipos de ítems (con respuesta binaria o de escala)

mediante la inclusión de una matriz de diseño. Para los de respuesta binaria se especificó el modelo logístico de un parámetro y para los de escala, el modelo de crédito parcial (Masters, 1982; Masters y Wright, 1997). Se llevan a cabo pruebas de ajuste de los ítems al modelo, que se basan en los residuos o diferencias entre las frecuencias observadas y las esperadas según el modelo, que se resumen en un índice (INFIT). Para el análisis de los ítems con el modelo de coeficientes mixtos se utilizó el programa *ConQuest* (Wu et al., 1997).

La estimación de los parámetros de los ítems (dificultad) se conoce en psicometría con el nombre de «calibración». El procedimiento utilizado permite obtener errores típicos de los parámetros e intervalos de confianza. Estos intervalos son importantes para determinar la adecuación de un ítem a un determinado país, ya que en PISA se eliminaron los ítems que en un país estuviesen fuera del intervalo de confianza. La calibración se llevó a cabo en muestras nacionales (de cada país) y en una muestra de calibración internacional final, formada por 500 sujetos extraídos al azar de los 30 países de la OCDE.

Otras evaluaciones como el TIMSS utilizan modelos con más parámetros, en la línea de Lord y del *Educational Testing Service*. No obstante, los procedimientos de estimación, determinación de la escala y obtención de las puntuaciones son similares.

Estimación de las puntuaciones: valores plausibles

Los tests educativos pueden tener dos propósitos fundamentales: 1) medir conocimientos y destrezas de estudiantes particulares, donde es muy importante minimizar el error de medida sobre cada estudiante, especialmente si se van a tomar decisiones sobre las puntuaciones ó 2) evaluar conocimientos y destrezas de la población, en las que lo más importante es la minimización de los errores en la población. Debido al diseño matricial de ítems utilizado, deben usarse complejos procedimientos para la estimación de las puntuaciones de los sujetos con datos incompletos. La metodología que utiliza PISA y otras evaluaciones internacionales fue desarrollada por Mislevy y utilizada por primera vez en el NAEP (Mislevy, 1991, Mislevy, Beaton, Kaplan y Sheehan, 1992; Mislevy, Jonson y Muraki, 1992; Mislevy y Shechan, 1989) y está basada en la teoría de la imputación de valores ausentes o perdidos de Rubin (1987). Las puntuaciones no son puntuaciones individuales y no sirven para el diagnóstico de los sujetos, sino solamente para la estimación de parámetros poblacionales consistentes.

El problema que hay que resolver es que cada sujeto responde solamente a un número limitado de ítems del test y es preciso estimar de algún modo cómo sería su comportamiento en el total de los ítems utilizados en la evaluación. Para ello, se predicen estos resultados utilizando las respuestas a los ítems que ha contestado y otras variables (denominadas de «condicionamiento», que se obtienen de los cuestionarios de contexto). En vez de predecir una única puntuación, se genera una distribución a *posteriori* de valores para cada sujeto con sus probabilidades asociadas (generalmente se asume que es una distribución normal). De

esta distribución se obtienen aleatoriamente cinco valores denominados «valores plausibles», porque son de la propia distribución de cada sujeto. Esto se hace para prevenir el sesgo que se produciría estimando la habilidad solamente a partir de un conjunto reducido de ítems del dominio. La selección de varios valores es necesaria para estimar la varianza error derivada de la imputación. Para la estimación de estos valores se requiere usar algún tipo de *software* específico. Estos valores son obtenidos también con el programa *ConQuest*. Existe un programa de libre distribución, AM, desarrollado por Miles y Cohen (2002), que se puede obtener de www.air.org.

El procedimiento usado en PISA para obtener los valores plausibles es similar a los de otras evaluaciones (NAEP y TIMSS) y consiste en utilizar para la imputación las respuestas del sujeto a los ítems del dominio que le fueron presentados, con los valores de los parámetros de la calibración nacional y un conjunto de «variables de condicionamiento». Hay un primer bloque común a todos los países formado por cinco variables (cuadernillo, sexo, ocupación del padre y de la madre y media en matemáticas de la escuela). Las restantes variables de condicionamiento se construyen a partir del cuestionario del alumno en el que todas las variables se codifican en códigos *dummy* y se reduce su dimensionalidad utilizando la técnica de los componentes principales, en un número que explique hasta el 95% de la varianza total, que puede variar entre países.

Los estadísticos poblacionales se estiman usando cada uno de los valores plausibles separadamente. El estadístico poblacional referido en los informes es el promedio de los estadísticos obtenidos con cada uno de los valores plausibles. La integración de la varianza de la imputación en la varianza de los estimadores de los parámetros se explica en el *Data Analysis Manual* (OCDE, 2005b; 2005c).

Las escalas de habilidad o competencia

La aplicación del modelo de TRI lleva a una escala continua de habilidad o competencia en cada dominio que permite estimar la posición de sujetos e ítems. Normalmente, la métrica está indeterminada y se estima en puntuaciones típicas (media 0, desviación típica 1), que mediante una transformación lineal se convierten a la métrica de PISA con media 500 y desviación típica 100. Cuando una puntuación del sujeto (o del grupo) se encuentra próxima a un punto de la escala, es más probable que sea capaz de contestar con éxito a los ítems que están en o por debajo del punto, pero menos probable que realice las tareas que están por encima. En la página 187 y 188 del *Technical Report* (OCDE, 2005a), pueden verse los «mapas de ítems» de los cuatro dominios escalados en «habilidad». Posteriormente, los ítems se reescalan a media 500 y desviación típica 100 y se presentan ordenados del más alto al más bajo, junto con una descripción de las tareas que implican, así como con su clasificación en cuanto a contenidos, demandas y contexto de utilización. El continuo en el que se sitúan todos los ítems se convierte en seis «niveles», que representan bandas o intervalos de puntuaciones y facilitan la interpretación. Las convenciones para establecer las nive-

les del continuo están claras y pueden consultarse en los *Technical Reports* de 2000 (Adams y Wu, 2002) o del 2003 (OCDE, 2005b). Examinando los ítems que se encuentran en los diferentes niveles, grupos de expertos en la materia desarrollan el significado de cada uno en cuanto a lo que los sujetos de dicho nivel conocen y pueden hacer. En las páginas 261 a 267 del *Technical Report* (OCDE, 2005a), se encuentran las descripciones de los seis niveles en la escala global de matemáticas y en cada una de las subescalas.

LOS CUESTIONARIOS DE CONTEXTO

Los cuestionarios de contexto suelen cumplir tres funciones dentro de las evaluaciones internacionales: 1) son usados para definir subgrupos de la población de examinados permitiendo introducir cualificadores a los resultados (género, etnia, nivel educativo de los padres, tipo de escuela…); 2) se examinan las relaciones de sus variables con los datos de rendimiento educativo y sus influencias sobre éste, y 3) sus datos se utilizan como variables de condicionamiento en la imputación y estimación de los valores plausibles. La mayor parte de las evaluaciones internacionales utilizan tres tipos de cuestionario: de escuela, de profesores y de alumnos. En PISA, se utilizan solamente cuestionarios de escuela y de alumnos, lo que algunos consideran una limitación; no obstante, teniendo en cuenta que sus objetivos no están ligados al currículo, podría resultar difícil la interpretación de los datos derivados de los profesores. En la web de la OCDE pueden obtenerse los cuestionarios utilizados. Algunas cuestiones de ambos cuestionarios se tratan como ítems únicos (p. ej., el sexo del alumnado o el tamaño de la escuela), pero la mayor parte fueron diseñadas para la medida de índices o de constructos latentes. Las medidas de índice son simples transformaciones o combinaciones lineales de algunas variables, mientras que las medidas de constructos son estimaciones de rasgos latentes derivadas mediante procedimientos de escalamiento. En general, las variables de los cuestionarios fueron tratadas con gran rigor. Las medidas de estatus ocupacional fueron trasladadas al Índice Internacional de Estatus Ocupacional (ISEI, Ganzeboom et al., 1992). Los niveles educativos de los padres fueron codificados en la clasificación ISCED y recodificadas posteriormente en años de escolarización. En general, las transformaciones son óptimas para el análisis de los datos.

Con el cuestionario de la escuela también se construyeron diversos índices de los que resultan novedosos los de selectividad de las escuelas, la cultura de la evaluación y los de gestión y niveles de autonomía.

Una buena parte de los ítems de ambos cuestionarios son los destinados a la medida de constructos latentes. Para su elaboración se utilizaron las técnicas psicométricas más adecuadas para analizar la dimensionalidad (análisis factoriales exploratorios y confirmatorios) y su estabilidad o consistencia entre países (con análisis confirmatorios multi-grupo), así como la estabilidad entre países de las relaciones entre los constructos. Como valores de ajuste se utilizaron los habituales en estos modelos (Kaplan, 2002). Una vez establecidas las dimensiones, los ítems componentes fueron calibrados con modelos politómicos de respuesta al ítem y las puntuacio-

nes de los estudiantes individuales fueron estimadas con procedimientos de máxima verosimilitud y transformadas a escalas de media 500 y desviación típica 100. La calibración internacional se realizó sobre la muestra internacional de 500 alumnos de los 30 países de la OCDE. Los constructos obtenidos mediante estos procedimientos fueron tanto de nivel del alumno como del centro.

EL DISEÑO MUESTRAL, PESOS DE MUESTREO Y VARIANZAS DE LOS ESTIMADORES

En un detallado manual de muestreo (OCDE 1999), se exponen todas las especificaciones para obtener muestras representativas de la población objetivo de alumnos de 15 años escolarizados. El diseño es el habitual estratificado de conglomerados en dos etapas, con muestras en la primera etapa proporcionales al tamaño. Las unidades primarias de muestreo son las escuelas y las secundarias los alumnos de cada escuela, que se seleccionaron (35) de forma aleatoria a partir de listas de sujetos elegibles en cada escuela. Las variables de estratificación fueron explícitas e implícitas, pudiendo variar entre países. En España, las explícitas fueron la titularidad del centro, comunidad autónoma, tamaño de la escuela y modalidad de enseñanza en el País Vasco; las implícitas, los códigos postales y en Cataluña el tamaño de la localidad. La selección efectiva de la muestra se lleva a cabo mediante un procedimiento sistemático con un intervalo de muestreo.

Las reglas de exclusión de escuelas y de estudiantes están completamente especificadas en el Manual. En todos los casos se determinó que la tasa de exclusión global intra-escuela debería ser inferior al 5%. Este resultado no se ha respetado completamente en todos los países, entre otros en España (7,9%). Se requirió una tasa de respuesta del 85% en las escuelas seleccionadas, que podría alcanzarse mediante escuelas de reemplazamiento (con ciertas limitaciones). La tasa de alumnos se fijó en el 80%. Las escuelas para la muestra de reemplazamiento fueron la anterior y la posterior a la escuela seleccionada, un interesante procedimiento que garantiza bastante bien la equivalencia de las escuelas sustitutas.

El tipo de muestreo llevado a cabo requiere de un proceso complejo de determinación de *pesos muestrales*, necesario antes de llevar a cabo los análisis estadísticos. Son varias las razones de la ponderación de los casos: afijación no proporcional en algunos estratos, falta de actualización de los marcos, ajustes derivados de la no respuesta de escuelas y alumnos dentro de las escuelas y recorte de pesos para prevenir influencias no deseadas de un pequeño conjunto de escuelas o estudiantes. Los procedimientos para la asignación de pesos en PISA reflejan los estándares de las buenas prácticas de las encuestas y son los mismos utilizados por la IEA y el NAEP. Una buena descripción puede encontrarse en el *Data Analysis Manual* (OCDE, 2005b, c).

El complejo diseño muestral lleva a dependencias entre las unidades o sujetos de la muestra (Kish, 1987) e implica procedimientos especiales para el cálculo de la varianza de los estimadores. En general, no existen fórmulas precisas, debiendo calcularse mediante métodos intensivos de cálculo, conocidos como de «re-

muestreo», que consisten en obtener múltiples muestras a partir de la muestra original. Rust y Rao (1996) destacan que el principio común de todos estos métodos es estimar el parámetro de interés para toda la muestra y en cada una de las muestras replicadas. La variabilidad entre las replicaciones resultantes se usa para estimar el error típico del estadístico bajo estudio. En la práctica, suelen utilizarse tres variantes: *jackknife* (utilizado en TIMSS), replicación repetida balanceada (BRR) y *bootstrap*. PISA utilizó el método BRR con la modificación de Fay (1989), que permite obtener más replicaciones (se obtuvieron 80) y mejora por tanto la precisión del estimador de la varianza. Una descripción extensa del procedimiento puede encontrarse en el *Data Analysis Manual* (OCDE, 2005b, c). En la varianza final de los estimadores, se incluye además el error debido al uso de los valores plausibles. Todos los cálculos se realizaron con el programa *WesVar* (Westat, 2000).

LOS ANÁLISIS DE TENDENCIAS Y LOS ENLACES DE TESTS

Los procedimientos de equiparación de puntuaciones por medio de ítems comunes antes mencionados son los que se utilizan en PISA-2003 para equiparar las puntuaciones de este estudio con las del 2000 en cada uno de los dominios. Estos procedimientos se implementan calibrando conjuntamente los ítems comunes a las dos evaluaciones. En matemáticas, se utilizaron 20 ítems comunes, en lectura 28 y en ciencias 25. Estos ítems comunes se denominan de enlace o «anclaje». En lectura y ciencias, se tomaron como referencia las medias del 2000 y en matemáticas y solución de problemas, las del 2003. No se establecieron tendencias para subdominios, ya que se requiere un mínimo de ítems para que funcione bien la equiparación. El subconjunto o selección de ítems comunes influye en la transformación, por lo que se producen errores de equiparación que deben ser tenidos en cuenta a la hora de establecer comparaciones entre las medias de 2000 y 2003. Normalmente, la presencia de ítems abiertos y codificados en escala complica el cálculo de los errores típicos (Kolen y Brennan, 2004). El manual técnico del PISA es muy escueto en lo que se refiere a los procedimientos de equiparación, a la calidad de los ítems seleccionados y al cálculo de los errores típicos, por lo que no podemos establecer juicios sobre la calidad de los datos de tendencias. La calidad de los datos de tendencias y las inferencias que se puedan extraer a partir de ellos depende de la semejanza en los contenidos de los tests, el ajuste de los ítems al modelo de respuesta al ítem, propiedades psicométricas de los ítems, tamaño de la muestra y de la posición de los ítems comunes en las pruebas que se equiparan.

PRESENTACIÓN DE LOS RESULTADOS

En el texto *Learning for Tomorrow's World* (OCDE, 2004), se presentan los resultados del dominio principal, los subdominios y los dominios secundarios. La forma de presentación de los resultados incluye la habitual mediante tablas

(las denominadas Tablas de Liga), que es la presentación preferida por los medios de comunicación y refuerza lo que algunos llaman el aspecto «Olympic Games» de las evaluaciones internacionales (Husen, 1987). No obstante, se introducen algunas novedades importantes. Se comienza con la presentación por países de los porcentajes de estudiantes en cada uno de los seis niveles, lo que proporciona una visión más adecuada de los resultados que las simples medias. Las tablas de medias vienen acompañadas de diagramas que representan la significación de las diferencias entre países, como en TIMSS. En las notas al capítulo, se presentan algunos cualificadores de los resultados y tamaños de efecto. Se presentan también para cada país las diferencias según el género, mostrando la significación y el sentido de las diferencias de forma gráfica y muy fácil de visualizar.

Las tendencias o cambios entre las medias de 2000 y 2003 se presentan de una forma muy clara acompañadas de los niveles de significación de las diferencias.

Las variables derivadas de los cuestionarios del alumno se presentan en porcentajes de respuesta en los dos niveles más altos de la escala para cada ítem, los valores medios de los índices tipificados y los coeficientes de la regresión o cambio en rendimiento por cambio unitario en la variable, así como el porcentaje de varianza del rendimiento explicada por la variable en el nivel de los resultados de los alumnos.

Se hace un interesante estudio del papel del contexto socioeconómico en la variación de los rendimientos y sus variaciones entre países, acompañados de diagramas de dispersión muy interesantes.

Finalmente, se analiza el ambiente del aprendizaje y la organización de la escuela por medio de análisis descriptivos y con técnicas de análisis «multi-nivel» (Raudenbush y Bryk, 2002), cada vez más habituales en los estudios de evaluación educativa, para establecer su influencia en la variación entre escuelas dentro de cada país, con los valores de los coeficientes en forma gráfica y los porcentajes de variación del rendimiento explicados por cada variable en cada país. Los análisis de la variación entre escuelas, que varía considerablemente entre países, son muy importantes para la reflexión sobre políticas educativas y consideraciones de equidad.

Finalmente, debemos destacar que los datos están disponibles para el análisis secundario por parte de otros investigadores en la web de la OCDE (www.pisa.oecd.org/pisa/outcome.htm). Aparecen muy detalladas todas las especificaciones de la base de datos. La disponibilidad de los datos para el análisis secundario también es habitual en las evaluaciones recientes de la IEA (TIMSS).

CONCLUSIONES

Para establecer una valoración crítica de la calidad metodológica de PISA-2003, se siguen los criterios de calidad establecidos para las evaluaciones internacionales por el *Board on International Comparative Studies in Education (BICSE)* (National Research Council, 1990):

- *Validez técnica del estudio*. La calidad técnica del estudio está en general fuera de dudas. PISA-2003 sigue todas las buenas prácticas de las evaluaciones internacionales en cuanto a representatividad de las muestras, calibración adecuada de los parámetros y de los ítems, estimaciones precisas de los resultados, buenas traducciones, adecuados cuestionarios de contexto bien codificados, y un plan adecuado de informes dirigidos a diversas audiencias. No obstante, se podrían plantear algunas cuestiones en ciertos aspectos que podrían mejorarse:
  - Aunque se ha cuidado mucho la validez de contenido y de constructo de las tareas de la evaluación, aún permanece por tratar un problema importante de las evaluaciones a gran escala, y es el de la «motivación de los alumnos» cuando participan en evaluaciones que no tienen consecuencias para ellos (*low-stakes*). Parece que los alumnos no se implican suficientemente en la resolución de las tareas (O'Neil, Abedi, Miyoshi y Mastergeorge, 2005; Wise y DeMars, 2003) y, como consecuencia, las puntuaciones de los tests podrían no ser indicadores válidos de lo que saben y pueden hacer. El problema puede ser más grave en las evaluaciones internacionales, dado que podrían producirse interacciones con las diferentes culturas de los países implicados. Deberían realizarse investigaciones en este sentido.
  - Piers (2003) ha planteado algunas objeciones sobre el control de las tasas de respuesta de diferentes países del estudio 2000, que se pueden trasladar a 2003 y no han sido adecuadamente respondidas por Adams (2003).
  - También se han planteado algunas dudas sobre la elección de la población objetivo (Piers, 2003) basada en la edad y no en el curso. Aunque no está claro en la literatura cuál es la mejor elección (Chromy, 2002), la cuestión no está resuelta. La selección por edad permite eliminar algunos problemas ligados a las diferentes políticas de repetición de los países, itinerarios formativos, etc., pero deja sin resolver otras cuestiones como las diferentes edades de entrada en la escuela que pueden afectar a los resultados.
  - La presentación de los resultados, muy cuidada, elaborada y de fácil comprensión, debería introducir con mayor visibilidad algunos cualificadores del rendimiento de los países: edad de entrada en la escuela, tasas de la población con educación secundaria, etc.

- *El estudio está ligado a trabajos previos para propósitos de comparación*. Las comparaciones longitudinales o tendencias son muy importantes para los países participantes, especialmente cuando como resultado de las eva-

luaciones se introducen cambios (Goldstein, 1995, 2004). El PISA está diseñado para proporcionar estos datos desde sus inicios y en 2003 presenta un estudio de tendencias. No obstante, como se ha señalado anteriormente, es la parte menos detallada del *Technical Report* y debería mejorarse su explicación.

- *El estudio tiene en cuenta las diferencias culturales entre países y se caracteriza por la neutralidad de la investigación.* La participación de los países en todas las fases elimina sesgos, pero no está claro que se haya logrado este objetivo, ya que, como señala Goldstein (2004), en la práctica el control real es ejercido por un pequeño grupo con responsabilidades ejecutivas. Por otra parte, PISA no tiene en cuenta los currículos nacionales por no considerarlos relevantes para los objetivos, pero se han encontrado algunas cuestiones difíciles para los alumnos de algunos países que llevan a algunos autores a pensar en la falta de neutralidad (Goldstein, Bonnet y Rocher, sometido a publicación y citado en Goldstein, 2004). Finalmente, PISA ha cuidado en general muy bien las variables contextuales importantes desde la perspectiva sociológica, pero aún quedan fuera algunos aspectos de índole psicológica y cultural que aparecen bien descritos en el trabajo de Bempechat, Jiménez y Boulay (2002), pero que son difíciles de hacer operativos.
- *El estudio tiene impacto sobre las políticas* y *prácticas educativas de los participantes.* Ésta es una cuestión que no podemos evaluar, puesto que no disponemos de informaciones de los países. En general, el estudio ha tenido una importante repercusión mediática, pero desconocemos su impacto en otros ámbitos. No obstante, la desvinculación del currículo sostenida en PISA, a diferencia de otros estudios como TIMSS, es posible que limite sus implicaciones en las políticas y prácticas educativas. El hecho de que los sujetos estén dispersos entre diferentes cursos o grados también puede dificultar este punto. Deberán hacerse más estudios sobre sus influencias en la política y prácticas educativas y su contribución a la comprensión de la calidad de la educación.

REFERENCIAS BIBLIOGRÁFICAS

ADAMS, R. J. (2003): «Response to "cautions on OECD's recent educational survey (PISA)"», en *Oxford Review of Education, 29,* pp. 38-389.

ADAMS, R. J.; WILSON, M. R; WANG, W. (1997): «The multidimensional random coefficients multinomial logit", en *Applied Psychological Measurement, 21*, pp. 1-24.

ADAMS, R. J.; WU, M. L. (2002): *PISA 2000 Technical Report.* Paris, OECD.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1999) : *Standards for Educational and Psychological Tests.* Washington DC, American Psychological Association.

BEATON, A. E.: «Item sampling», en KEEVES, J. P. (ed.). (1997): *Educational research, methodology, and measurement: An international handbook* (2nd.ed.) (pp. 976-984). Cambridge, Pergamon.

BEATON, A. E.; POSTLETHWAITE, T. N.; ROSS, K. N.; SPEARRITT, D.; WOLF, R. M. (1999): *The benefits and limitations of international educational achievement studies.* Paris, International Institute for Educational Planning, UNESCO.

BEMPECHAT, J.; JIMÉNEZ, N. V.; BOULAY, B. A. (2002): «Cultural cognitive issues in academic achievement: New directions for cross-national research», en PORTER, A. C.; GAMORAN, A. (eds.): *Methodological advances in cross-national surveys of educational achievement.* Washington DC, National Academy Press.

BRENNAN, R. L. (2002): *Elements of Generalizability Theory*. New York , Springer.

CHILDS, R; JACIW, A. P. (2003): «Matrix sampling of items in large-scale assessments», en *Practical Assessment, Research, and Evaluation, 8*, pp. 16-28.

CHROMY, J. R. (2002).: «Sampling issues in design, conduct, and interpretation of international comparative studies of school achievement», en PORTER, A. C.; GAMORAM, A. (eds.): *Methodological advances in cross-national surveys of educational achievement.* (pp. 80-114). Washington DC, National Academy Press.

CRONBACH, L. J.; GLESER, G. C.; NANDA, N; RAJARATNAM, N. (1972): *The Dependability of Behavioral measures: The generalizability theory for Scores and Profiles*. New York, Wiley.

EUROPEAN COMMISSION (2002): *Education and training in Europe: Diverse systems, shared goals for 2010*. Luxemburg, EC.

— (2004): *New indicators on education and training.* Brussels, Commission Staff Working Paper.

FAY, R. E. (1989): «Theoretical application of weighting for variance calculation», en *Proceedings of the Section on Survey Research Methods of the American Statistical Associatio,* pp. 212-217.

FITPATRICK, A. R.; LEE, G; GAO, F. (2001): «Assessing the comparability of school scores across test forms that are not parallel», en *Applied Measurement in Education, 14,* pp. 285-306.

FLODEN, R. E. (2002): «The measurement of opportunity to learn», en PORTER, A. C.; A. GAMORAM, A. (eds.): *Methodological advances in cross-national surveys of educational achievement.* (pp. 231-266). Washington DC, National Academy Press.

GANZEBOOM, H. B. G.; GRAAF, P. M. De; TREIMAN, D. J. (1992): «A standard international socioeconomic index of occupational status», en *Social Science Research, 21,* p. 56.

GIFI, A. (1990): *Nonlinear Multivariate Analysis.* New York, Wiley.

GOLDSTEIN, H. (1995): «Interpreting international comparisons of student achievement», en *Educational Studies and Documents.* París, UNESCO.

— (2004): «The education world cup: International comparisons of student achievement», en *Plenary Talk to Association for Educational Assessment-Europe.* Budapest, noviembre de 2004.

GREANEY, V.; KELLAGHAN, T. (1996): *Monitoring the learning outcomes of education systems.* Washington DC, The World Bank.

GREGORY, K. D; MARTIN, M. O. (2001): *Technical standards for IEA studies: An annotated bibliography.* Amsterdam, IEA.

HAMBLETON, R. K. (2002): «Adapting achievement tests into multiples languages for international assessments», en PORTER, A. C.; GAMORAM, A. (eds.): *Methodological advances in cross-national surveys of educational achievement.* (pp. 58-79). Washington DC, National Academy Press.

HAMBLETON, R. K.; MERENDA, P. F.; SPIELBERGER, C. D. (eds.)(2005): *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah NJ, Erlbaum.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, J. (1991): *Fundamentals of Item Response Theory.* Newbury Park CA, Sage.

HUSEN, T. (1987): «Policy impact of IEA research», en *Comparative Education Review, 31* , pp. 129-136.

INTERNATIONAL LABOUR ORGANIZATION (1990): *International Standard Classification of Occupations: ISCO-88.* Ginebra, International Labour Office.

KAPLAN, D. (2000): *Structural equation modelling: Foundation and extension.* Thousand Oak CA, Sage.

KEEVES, J. P. (1997): «Validity of tests», en KEEVES, J. P. (ed.): *Educational research, methodology, and measurement: An international handbook* (2nd.ed.) (pp. 976-984). Cambridge UK, Pergamon.

KISH, L. (1987): *Statistical design for research.* New York, Wiley.

KOLE, M. J.; BRENNAN, R. L. (2004): *Test equating: Methods and Practices.* New York, Springer.

LEVIN, H.: «Educational performance standards and the economy», en *Educational Researcher, 27(4),* 1988, pp. 4-10.

LINN, R. L. (1993): «Linking results of distinct assessments», en *Applied Measurement in Education, 6,* pp. 83-102.

— (2002) «The measurement of student achievement in international studies», en PORTER, A. C.; GAMORAM, A. (eds.) : *Methodological advances in cross-national surveys of educational achievement.* (pp. 27-57). Washington DC, National Academy Press.

LORD, F. M. (1980): *Applications of item response theory to practical testing problems.* Hillsdale NJ, Erlbaum.

MARTIN, M. O.; GREGORY, K. D.; STEMLER, S. E. (2000): *Third International Mathematics and Science Study: 1999. Technical Report.* Chestnut Hill MA, Boston College.

MARTIN, M. O.; MULLIS, I. V. S.; CHROSTOWSKI, S. J. (2004): *TIMSS 2003. Technical Report.* Chestnut Hill MA, Boston College.

MARTIN, M. O.; RUST, K.; ADAMS, R. J. (1999): *Technical standards for IEA studies.* Ámsterdam, IEA.

MARTÍNEZ ARIAS, R. (1995): *Psicometría: Teoría de los tests psicológicos y educativos.* Madrid, Síntesis.

MASTERS, G. N. (1982): «A Rasch model for partial credit scoring», en *Psychometrika, 47*, pp. 149-174.

MASTERS, G. N.; WRIGHT, B. D. (1997): «The Partial Credit Model», en VAN DER LINDEN, W. J.; HAMBLETON, R. K. (eds.) (1997): *Handbook of Modern Item Response Theory.* New York, Springer.

MILES, J.; COHEN, J. (2002): *AM Statistical Software.* Washington DC, American Institutes for Research.

MILLMAN, J.; GREENE, J. (1989): «The specification and development of tests of achievement and ability», en LINN, R. L. (ed.): *Educational Measurement* (3rd ed.) (pp.335-366). New York, Macmillan, .

MISLEVY; R. J.; STEINBERG, L. S.; ALMOND, R. G. (2003): «On the structure of educational assessments», en *Measurement: Interdisciplinary Research and Perspectives, 1*, pp. 3-67.

MISLEVY, R. J. (1991): «Randomization-based inference about latent variable from complex samples», en *Psychometrika, 56*, pp. 177-190.

— (1992) *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton NJ, Educational Testing Service.

— (1995) «What can we learn from international assessments?», en *Educational Evaluation and Policy Analysis, 17*, pp. 410-437.

MISLEVY, R. J.; SHEEHAN, K. M. (1989): «The role of collateral information about examinees in item parameter estimation», en *Psychometrika, 54*, pp. 661-679.

MISLEVY, R. J.; BEATON, B.; KAPLAN; SHEEHAN, K. M. (1992): «Estimating population characteristics from sparse matrix samples of item responses», en *Journal of Educational Measurement, 29*, pp. 133-161.

MISLEVY, R. K.; JOHNSON, E. G.; MURAKI, E. (1992): «Scaling procedures in NAEP», en *Journal of Educational Statistics, 17*, pp. 131-154.

NATIONAL RESEARCH COUNCIL (1990): *A framework and principles for international comparative studies in education.* Board on International Comparative Studies in Education. Washington DC, National Academy Press.

NISHISATO, S. (1980): *Analysis of Categorical Data: Dual Scaling and its Applications.* University of Toronto, .

O'NEIL, H. F.; ABEDI, J.; MIYOSHI, J. (2005): MASTERGEORGE, A.: «Monetary incentives for low-stakes tests», en *Educational Assessment, 19* pp. 185-208.

OCDE (1989): *Employment outlook*. París, OCDE.

— (1996)*Lifelong learning for all*. París, OCDE.

— (1996) *PISA Sampling Manual, main study*. París, OCDE.

— (2001) *Knowledge and Skills for Life: First Results from PISA 2000*. París, OCDE.

— *(2003) The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving.* París, OCDE.

— (2004) *Learning for Tomorrow's World – First Results from PISA 2003.* París, OCDE.

— (2005a) *PISA 2003. Technical Report.* París, OCDE.

— (2005b) *PISA 2003 Data Analysis Manual: SAS® Users.* París, OCDE.

— (2005c) *PISA 2003 Data Analysis Manual: SPSS® Users.* París, OCDE.

PORTER, A. C.; GAMORAN, A. (eds.) (2002): *Methodological advances in cross-national surveys of educational achievement.* Washington DC, National Academy Press, .

POSTLETHWAITE, T. N. (1999): *International studies of educational achievement: Methodological issues.* Hong Kong. University of Hong Kong, Comparative Education Research Center.

PRAIS, S. J. (2003): «Cautions on OECD's recent educational survey (PISA)», en *Oxford Review of Education, 29*, pp. 139-163.

RASCH, G. (1980): *Probabilistic models for some intelligence and attainment tests.* Copenhague, Nielsen and Lydiche, 1960-1980 (Reeditado en 1980 por University of Chicago Press).

RAUDENBUSH, S. W.; BRYK, A. S. (2002): *Hierarchical Linear models in social and behavioral research.* Thousand Oaks CA, Sage.

ROBINSON, P. (1999): «The tyranny of league tables», en ALEXANDER, R.; BROADFOOT, P.; PHILLIPS, D. (eds.): *Learning from comparing-new directions in comparative educational research. Vol. 1.* London, Symposium Books.

RUBIN, D. B. (1987): *Multiple imputation for nonresponse in surveys.* New York, Wiley.

RUST, K.; RAO, J. N. K. (1996): «Variance estimation for complex surveys using replication techniques», en *Statistical Methods in Medical Research 5* , pp. 283-310.

THISSEN, D.; WAINER, H. (2001): *Test scoring.* Mahwah, Erlbaum.

THISSEN, D.; ORLANDO, M. (2001): «Item response theory for items scored in two categories», en THISSEN, D.; WAINER, H. (eds.): *Test scoring* (pp. 73-140). Mahwah, Erlbaum.

VAN DER LINDEN, W. J.; HAMBLETON, R. K. (eds.) (1997): *Handbook of Modern Item Response Theory.* New York, Springer.

WESTAT (2000): *WesVar complex samples. 4.0.* Rockville MD, Westat.

WISE, S. L.; DEMARS, C. E. (2003): «Examinee motivation in low-stakes assessment: Problems and potential solutions», en *Paper presentd at the Annua l Meeting of the American Association of Higher Education Assessment Conference.* Seattle, June.

WU, M. L.; ADAMS, R. J.; WILSON, M. R. (1997): *ConQuest: Multiaspect test software* [computer program]. Camberwell, Australia, Australian Council for Education Research.