Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Una crítica metodológica de las evaluaciones PISA

A Methodological Critique of the PISA Evaluations

Fernandez-Cano, Antonio
University of Granada (Spain)

Abstract

This paper conducts a methodological evaluation of the PISA international evaluations, giving a critical analysis of their shortcomings and limitations. A methodological review or meta-evaluation has been carried out on the multiple PISA reports in an attempt to demonstrate the plausible validity of the inferences that PISA maintains given a series of methodological limitations such as: an inconsistent rationale, opaque sampling, unstable evaluative design, measuring instruments of questionable validity, opportunistic use of scores transformed by standardization, reverential confidence in statistical significance, an absence of substantively significant statistics centered on the magnitudes of effects, a problematic presentation of findings and questionable implications drawn from the findings for educational norms and practice. There is an onus on PISA to provide and demonstrate more methodological rigor in future technical reports and a consequent need to be show great caution lest unfounded inferences are drawn from their findings.

Reception Date
2016 January 21

Approval Date
2016 May 10

Publication Date:
2016 May 11

Keywords:

Evaluation; Meta-evaluation; Methodology of evaluation; PISA

Resumen

En este trabajo realizamos una evaluación metodológica de las evaluaciones internacionales PISA, presentando un análisis crítico de sus deficiencias y limitaciones. Presentamos una revisión metodológica o meta-evaluación de los múltiples informes PISA, en un intento de demostrar la validez plausible de las inferencias que PISA mantiene, teniendo en cuenta una serie de limitaciones metodológicas tales como: una lógica incoherente, toma de muestras opacas, diseño evaluativo inestable, instrumentos de medición de validez cuestionables, el uso oportunista de las puntuaciones transformadas por la normalización, la confianza reverencial en la significación estadística, la ausencia de estadísticas sustantivamente importantes centradas en las magnitudes de los efectos, una presentación problemática de los hallazgos e implicaciones cuestionables extraídas de los resultados para las prácticas y las legislaciones educativas. Recae sobre PISA la responsabilidad de proporcionar y demostrar mayor rigor metodológico en los futuros informes técnicos y la consiguiente necesidad de ser cuidadosos para no mostrar inferencias sin fundamento a partir de sus hallazgos.

Fecha de recepción
21 Enero 2016

Fecha de aprobació
10 Mayo 2016

Fecha de publicació
11 Mayo 2016

Palabras clave:

Evaluación; Meta-evaluación; Metodología of evaluación; PISA.

PISA (*Programme for International Student Assessment*: Programa para la valoración internacional del alumno) es un estudio comparativo estandarizado relativo a evaluaciones internacionales a gran escala, que está siendo desarrollado en un serie de países participantes, miembros o asociados a la OCDE (Organización para la Cooperación y Desarrollo Económico: *Organization for Economic Co-operation and Development*). Como estudio a gran escala, PISA puede tener implicaciones sobre una serie de afectados; así pues, es importante revisar la información que el estudio ofrece con una perspectiva crítica.

Corresponding author / Autor de contacto

Fernandez-Cano, Antonio. Universidad de Granada. Facultad de Ciencias de la Educación. Campus Universitario La Cartuja. 18071-Granada (España). (afcano@ugr.es).

PISA se implementa desde el año 2000 y cada tres años para evaluar tres competencias escolares básicas: comprensión lectora (en formatos de papel y digital), competencia matemática y competencia científica; aunque en cada trienio se enfatice un determinado dominio competencial. La evaluación de 2012 fue administrada a 510.000 alumnos de 15 años, con un grado escolar que variaba según el país participante representando a 28 millones de jóvenes de 65 países (OECD, 2013a).

Los archivos de datos para su tratamiento analítico que implican a 65 países distinguen en PISA 2012 varios grupos de variables relativas al alumno (470) y variables relativas al centro/escuela (249), pero no se consideran variables profundamente pedagógicas acerca de las mejores prácticas de una enseñanza eficiente para el aprendizaje escolar; sólo a partir de PISA 2012, se empezaron a considerar estas cuestiones en una sección titulada "Gobernanza escolar, currículo y valoración" (OECD, 2015a).

PISA está en la línea de una tradición de evaluaciones educativas internacionales cuyos antecedentes serían los estudios IEA (*International Educational Achievement*) realizados desde 1958 por *International Association for the Evaluation of Educational Achievement* (IEA, 2012) y cuyos informes evaluativos más recientes y relevantes son: TIMSS (*Trends in International Maths and Science Study*) (ver Mullis, Martin, Minnich, Stanco, Arora, Centurino & Castle, 2012) y PIRLS (*Progress in International Reading Literacy Study*) (ver Mullis, Martin, Minnich, Drucker & Ragan, 2012). Tanto TIMSS como PIRLS difieren de PISA en que aquellos enfatizan la dimensión curricular de las prácticas en el aula, mientras PISA enfatiza el aprendizaje al margen de currículo escolar que permite a los estudiantes aplicar procesos y contenidos al contexto del mundo real. En EE.UU. se viene implementando desde 1964 el programa NAEP (*National Assessment of Educational Progress*) con información disponible en *National Center for Education Statistics* (2012).

La meta-evaluación de este programa de acuerdo con las orientaciones de Scriven es una necesidad perentoria para su mejora, interpretación, difusión y un uso procedente de sus hallazgos (Berliner, 2011; Scriven, 2011). Empero, la consistencia metodológica de las evaluaciones PISA presenta aspectos cuestionables metodológicamente y la teoría del programa debiera aún probar la validez de sus supuestos subyacentes sobre todo la validez interna, que algunos (e. g. Hanberger, 2014) consideran baja.

A la luz de los criterios del *American Joint Committee on Standards for Educational Evaluation* (Yarbrough, Shulha, Hopson & Caruthers, 2011), se exponen una serie de cuestiones sobre estándares de exactitud buscando asegurar que una evaluación como PISA revelará y convergerá información técnica (metodológicamente) adecuada. Debemos recordar que los estándares de exactitud pretenden incrementar la dependencia y valía de las representaciones, proposiciones y hallazgos, especialmente de aquellos que sostienen interpretaciones y juicios sobre calidad.

Objetivo de este estudio

Este estudio no pretende realizar una crítica demoledora de PISA; antes bien, sólo tiene una pretensión "metaevaluativa" de un programa de evaluación, cual es PISA, y que constituye un esfuerzo ingente, loable pero mejorable, con miles de agentes participantes, millones de alumnos evaluables y un ingente gasto económico. Así pues, este artículo no es una mera crítica a una encuesta sino sobre cómo PISA podría mejorarse.

El objetivo central de este estudio es doble: realizar en concreto un análisis metodológico de las evaluaciones PISA y en general de su método de encuesta a gran escala asociado; revisando bastantes tópicos afines a PISA para elaborar un argumento efectivo: su mejorable metodología de la investigación. PISA no puede ser el referente arquimediano y juez único de la evaluación de sistemas educativos y por tanto sus resultados e indicadores debieran interpretarse dentro de un contexto cultural concreto y un marco

conceptual preciso, cual es la idea de logro vital-funcional o competencia (Meyer & Benavot, 2013).

La metodología aquí utilizada es el análisis crítico metodológico o meta-evaluación (Scriven, 2011; Stufflebeam, 2011). Estudios meta-evaluativos como éste debieran ser una práctica habitual y deseable aunque hay que reconocer su dificultad y el gran riesgo inherente en toda meta-evaluación como una amenaza a su validez: que quede reducida a una sarta de enunciados desconectados sin relación algún con la cuestión que se evalúa. Además, existe otra amenaza, relacionada con la falta de exactitud metodológica y cobertura inadecuada de información importante acerca de la meta-evaluación de programas internacionales, que detectó Bollen, Paxton y Morishima (2005).

Concretando, este texto considera una serie de limitaciones metodológicas de PISA que impactan las inferencias que puede extraerse de los resultados.

Revisando la aproximación metodológica de *PISA*

Versiones oficiales de la metodología de PISA están disponibles en Adams (2003); Adams, Berezner y Jakubowski (2010); Adams, Wu y Castensen (2007), y sobre todo en los sucesivos informes técnicos de la OCDE, especialmente el informe técnico de PISA 2009 (OECD, 2012).

Las primeras críticas metodológicas llegaron pronto, de parte de Prais (2003) para datos de PISA 2000 en Gran Bretaña. Prais (2003) cuestionaba el constructo subyacente de los instrumentos, por ajeno a los currículos escolares, el uso de muestras improcedentes, ya que incluían a alumnos mayores de 15 años, sobre todo repetidores, y tasas de respuesta muy bajas (del 60 %). Prais sugería que insuficiencias metodológicas en PISA habían producido una aparente mejora del rendimiento de los estudiantes británicos, particularmente cuando se comparaba con sus compañeros suizos y alemanes

Pronto fue tildado de ignorante de la metodología de los estudios evaluativos internacionales y de la de PISA en particular por Adams (2003), coordinador de PISA 2000.

Brown, Micklewright, Schnepf y Waldmann (2007) cuestionaban la robustez de los hallazgos de PISA y otros estudios evaluativos internacionales (TIMSS y PIRLS) y la procedencia de generar puntuaciones agregadas a partir de las respuestas de los respondientes. Dreschsel, Carstensen y Prenzel (2011) tratan problemas metodológicos y describen requisitos de la construcción de tests para futuras evaluaciones a la luz de su significación en la educación científica.

PISA y en general los estudios comparados internacionales son genéricamente criticables sobre todo en la metodología de encuesta que utilizan, en la cuestionable validez de los instrumentos de medida que utilizan por su escasa relación con los currículos nacionales (Dolin & Krogh (2010) y en serias pseudo-interpretaciones de los resultados entre agentes no cualificados (Carnoy & Rothstein, 2013).

Cuestionamientos metodológicos

A continuación se exponen pormenorizadamente una serie pormenorizada de cuestionamientos metodológicos, que nos advierten de las limitaciones plausibles de estas evaluaciones PISA, acerca del estándar general de exactitud.

Racionalidad inconsistente

PISA como evaluación externa sumativa está plena de inconsistencias que raya la irracionalidad de este proceso; especialmente debiera advertirse que la OCDE no impone PISA sino que son los ministerios de educación los que eligen participar aunque el modelo evaluativo sí sea impuesto, no demandado, ni deseado, ni consentido, ya que es una imposición política gubernamental sin ninguna consideración a las opiniones de alumnos, padres y profesores.

Las hipótesis correlacionales subyacentes están mal enunciadas y sus sucesivas concatenaciones adolecen de fundamento

(Banks, 2012; De Witte & Kortelainen, 2013). PISA hipotetiza que: "La riqueza económica correlaciona positivamente con el dominio de conocimientos y competencias de sus escolares", que "el rendimiento escolar predicto sobre la base de ciertas variables académicas depende de ciertas características de las escuelas"; en consecuencia, si las escuelas de un sistema escolar verifican ciertas características facilitadores del rendimiento, la riqueza de ese país se verá mejorada. Tanto la hipótesis inicial como las secuenciales son un tanto simplistas. Este supuesto pudiera haberse mantenido en base a un hallazgo temprano procedente de PISA 2003 estimado por una correlación promedio de 0.32 (que explicaba alrededor del 10 % de la varianza de las puntuaciones de rendimiento) pero con considerables variaciones según países participantes.

Los factores explicativos que PISA propone como relevantes son de naturaleza correlacional y no causal; en consecuencia, ninguna decisión causal puede generalizarse como subrepticiamente hace PISA utilizando modelos de regresión múltiple o modelo de multinivel. Estos modelos son estadísticos y no admiten relaciones causales por mucho que los hallazgos de PISA obtenidos y expuestos en los propios informes sean erradamente interpretados como inferencias causales *ad nauseam*; una relación sobre advertencias acerca de inferencias en PISA 2009 puede leerse en Ercikan, Roth y Asil (2015). Concretamente su inferencias sobre los que hace que las escuelas sean exitosas (OECD, 2015b; Vol. IV); la relación entre alumnos y dinero-gasto (OECD, 2015b; Vol. VI); la causación espuria por única y lineal de que la predisposición para el aprendizaje es condicionada por el compromiso, dirección y auto-creencias del alumno (OECD, 2015b; Vol. III).

PISA es una evaluación que no tiene consecuencias directas para la promoción de la persona evaluada (estudiante de 15 años), de aquí que el posible bajo nivel de motivación y por tanto de implicación del examinando en las tareas. Un índice del uso de PISA para la valoración personal de alumno fue derivado de ocho ítems preguntadas a directores escolares, pero sólo un de eso ítems se usó con el propósito de tomar decisiones sobre promoción o retención de alumnos (OECD, 2013b) oscilando desde un 1% en Islandia a un 98 % en Portugal (OECD, 2013b, p. 149). Su propósito no es la mejora de los alumnos o del sistema nacional de educación, en el caso español (Couso, 2009), en el caso ruso (Zuckerman, Kovaleva & Kuznetsova, 2013), en Taiwan (Zhang & Sheu, 2013) o en Alemania con los alumnos que repiten en matemáticas (Ehmke, Drechsel & Carstensen, 2008); antes bien, se trata de un sutil ejercicio de rendir cuentas y competitividad tendente a justificar el gasto económico del momento o racionalizar el gasto futuro en educación.

Es difícil sumir que PISA sea un modelo unificado basado en competencias para situaciones de la vida real (DeSeCo, 2008; OECD, 2013a, p. 3; OECD, 2014a, p. 4; Rychen & Salganik, 2003) ya que los desempeños vitales de una campesina egipcia, un pasto andino, una ejecutiva de Tokio o un obrero industrial alemán *no* son los mismos. Además, las competencias en la estructuras de PISA han estado sujetas a considerable evolución a través de las diversas encuestas PISA realizadas hasta ahora.

Muestreo opaco

PISA suele aportar una extensa documentación técnica en su múltiples informes acerca de tasas de exclusión y procedimientos de muestreo; particularmente los tamaños muestrales son cuidadosamente expuestos y descritos. Sin embargo, ciertas omisiones sobre el muestreo quedan patentes: se adopta cuestionablemente el supuesto de hay grupos comparables de alumnos en base a la respectivas muestras nacionales, pero no son estrategias de muestreo basadas en la edad ni en el grado las que permitan alcanzar muestras equilibradas en término de edad y tiempo de escolarización (Strietholt, Rosén & Bos, 2013).

Krohne, Meier y Tillmann (2004) habían denunciado para el caso alemán la exclusión

de alumnos con necesidades educativas especiales, que eran evaluados diferencialmente o excluidos, la inclusión de repetidores y la diversidad de itinerarios (*tracks*) suscitan interrogantes sobre la idoneidad de la población encuestada. Estas restricciones hacen cuestionable que PISA sea un programa inclusivo.

Los saltos muestrales no se suelen exponer claramente tanto en técnica de muestreo utilizada, tamaños muestrales, errores de muestreo asociados y sobre todo porcentajes de mortalidad ante bajas tasas de retorno, en torno al 15 % en PISA 2006, del 25 % en PISA 2009 y del 20 % en PISA 2012; tasa de excluidos, que no debiera superar el 5 % de los alumnos seleccionados de la población-diana, o peor aún que existan tasas diferenciales de exclusión según países (mortalidad diferencial). Además, hay países en los que la escolaridad ya no es obligatoria a los 15 años, como PISA rígidamente insistir en adoptar. El muestreo se torna más opaco pues no hay constancia de qué acuerdos internacionales se adoptan ante una alta mortalidad que pudiera ser crítica.

Otro aspecto preocupante en el muestreo es el tamaño muestral de los países. No deja de ser cuestionable la diversidad de tamaños no proporcionales a la población del país por mucha florituras estadísticas a que se apela; por ejemplo, se hace difícil admitir la comparabilidad de un país como Austria, con apenas 8 millones habitantes, que cuente en PISA 2009 con una muestra total de 6.590 estudiantes sobre una población accesible de 82.135, frente a Japón, con 126 millones sea representado por una muestra de 6.088 alumnos para una población accesible de 1.060.381 (OECD, 2012, p. 188). Entonces, Knipprath (2010) clamaba que la calidad de la educación japonesa no había sido bien investigada a causa de una muestra no representativa con un conjunto corto de datos. Por otro lado y como rasgo positivo, algunas muestras naciones están sobre-muestreadas, por ejemplo en 2013 países como Bélgica, Colombia, Italia y especialmente España invirtieron en sobre-muestreo para obtener datos a nivel regional.

PISA confía en el uso de re-muestreo[2], en concreto a la replicación repetida equilibrada también conocida por las siglas en inglés BBR (*Balanced Repeated Replication*) usando el método de Fay (Judkins, 1990); una técnica estadística usada para generar estimaciones de los errores típicos específicos de cada país y obviamente no un re-muestreo de los alumnos tomados individualmente; pero recuérdese que si la muestra inicial no es representativa no tiene sentido apelar al re-muestreo, ni a técnica de replicabilidad alguna (dígase *bootstrap*[3], *jacknife*, *BRR* o validación cruzada).

Endeble diseño evaluativo

El diseño de PISA es muy simple, básicamente un diseño observacional de tendencias, que no un deseable diseño longitudinal de panel o cohortes, como el ampliamente utilizado en su programa análogo NAEP (*National Center for Education Statistics*, 2012). PISA aún está un tanto lejos de NAEP[4] ante todo a nivel metodológico ya que NAEP usa diseños longitudinales.

La estructura de medida para evaluar la efectividad educativa de PISA continúa confiando sucesivamente en un único indicador (desempeño de muestras grandes en un test a gran escala) y en el uso de estas puntaciones en el test aisladas de otros indicadores que también captan lo que significa ser efectivo. En esta línea de pluralismo métrico, un estudio longitudinal consistente con todas las aplicaciones tri-anuales de PISA serían indudablemente bienvenido y fácil de realizar ya que los datos primarios de PISA están disponibles en los diversos informes PISA (OECD, 2009; OECD, 2013a), pero con la reserva de que los instrumentos, que se cambian de una edición a otra, no son iguales ni paralelos ya que algunos ítems se descartan, otros nuevos se introducen y algunos anteriores se enmiendan; no obstante PISA dispone de bancos de ítem calibrados. Es manifiesta la cuestión de los errores de anclaje (*linking*) entre ítems, ya

reconocida y tratada en PISA 2009 (OECD, 2012, p. 143-146); hablamos de discrepancias en los índices de dificultad (porcentaje internacional correcto, para usar la expresión de PISA) entre una aplicación y otra (OECD, 2012, p. 215-228).

La sarta de inferencias causales, que de PISA se han derivado, debiera interpretarse con múltiples reservas, las propias de todo diseño observacional con técnicas correlacionales de análisis de datos, incluidos pesos de regresión, por muy sofisticadas que éstas sean. Recuérdese que el venerable *dictum* metodológico: "Nunca correlación es sinónimo de causa" siempre es pertinente, aunque PISA se curé en salud un tanto tardíamente y casi a hurtadillas (OECD, 2013, p. 29) al enunciar que: "PISA no mide causa y efecto"; pero bastante de las inferencias que ha posibilitado han sido espuriamente causales.

PISA omite variables "objetivas" moderadoras sólo incluibles tras un diseño experimental factorial; entonces, sí se podría hablar con propiedad de factores en una aproximación al modelado multinivel y no tan apresuradamente como hace PISA. También desconsidera covariantes incorporables como puntuaciones de propensión, manifiestamente significativas como gasto educativo, ratio alumno-profesor y retribución-supervisión del docente u otras más cualitativas pero igualmente objetivas como capacidad del alumno, tradición de plena escolarización o cualificación del profesor. En general, PISA incorpora variables sociológicas pero omite variables psicológicas, culturales y sobre todo pedagógicas, como las antes señaladas, que podrían ser relevantes según atestigua la investigación previa (Wang, Haertel &Walberg, 1993).

Cuestionable validez de los instrumentos de medida

Las pruebas de rendimiento propias de evaluaciones a gran escala presentan múltiples carencias relativas sobre todo a la validez, tales como:

Aplicaciones inconvincentes del modelo Rasch

En PISA la baja validez de las medidas podrían agravarse dada la naturaleza matricial de los ítems administrados, sobre los que no hay garantía de equivalencia, pues a cada examinando no se le administran todos los ítems y, en consecuencia, las comparaciones no están basadas en un test común sino que alumnos diferentes responden a cuestiones diferentes ya que no sucede que todo ítem del conjunto total (*pool*) de ítems aplicables sea administrado a todos los examinandos. No hay garantía entonces de que ni los cuadernillos, ni los ítems administrados sean equivalentes entre sí, lo cual plantea retos realmente insuperables para estimar los logros individuales.

Es un reduccionismo cuestionable considerar que los ítems incluidos en un cuadernillo son ítems con índices de dificultad conocidos usando solo estimaciones previas a la admiración. También es una consideración reduccionista aceptar como apropiado el uso del muestreo matricial de ítems solo porque sea bien conocido y ampliamente usado como la aproximación más común en evaluaciones a gran escala para estimar rendimiento poblacionales; pues como Rutkowsky (2014) aconseja este modelo de imputación (más comúnmente llamado un modelo condicionante) usado en datos PISA asume medidas completas, sin error, pues desviaciones de este supuesto puedan tener un impacto significativo sobre las estimaciones de parámetros del modelo condicionante, en estimaciones sub-poblacionales del rendimiento y en supra- o sobre-estimaciones de diferencias sub-poblacionales.

La validez de contenido de los instrumentos evaluativos es también harto cuestionable dado, por un lado, el reducido número de ítems que cada sujeto realiza sobre el libreto administrado al estudiante, y generados por muestreo matricial, con lo que no hay garantía y evidencia de que las formas paralelas de cada prueba sean en verdad paralelas y más sin disponer de índices de

discriminación de los ítems, que no se aportan en los informes.

Más cuestionable es apelar a procedimientos psicométricos de puntuación por evaluaciones adaptativas y confiar en al uso de la TRI (Teoría de Respuesta al Ítem) para generar formas paralelas, en concreto en al modelo Rasch, para equiparar puntuaciones procedentes de diversas formas del test, basado sólo en índices de dificultad pero sin captar todas las dimensiones relevantes del test; ello podría ser contraproducente. Kreiner y Christiansen (2014) aportan profunda evidencia de que el escalamiento utilizado en PISA en base al modelo Rasch es ampliamente inadecuado dado el alto funcionamiento diferencial de los ítems que socavaría la robustez de los escalafonamientos (*rankings*) de países.

Hubo además países que administraron ítems más sencillos para estudiantes de baja capacidad e incluso ciertos países desconsideraron ítems en base a "pobres propiedades psicométricas" pero que funcionaban bien en la vasta mayoría del resto de países (OECD, 2012, p. 132). Por otro lado, surge la cuestionable idoneidad de los ítems por escasa e injustificada relación con los currículos nacionales[5]; o menos aún con los libros de texto aprobados oficialmente para representar a tales currículos.

Los ítems a incluir también debieran disponer de su índice de discriminación. No es técnica ni estadísticamente imposible tener en cuenta el índice de discriminación. Obviamente, el modelo Rasch usando en PISA no debiera aplicarse entonces ya que implica a un único parámetro basado en el índice de dificultad calculado por el número de respuestas correctas. Un modelo de dos parámetros debiera considerarse a pesar de que las plausibles interpretaciones de resultados sean más difíciles de hacer. Por ejemplo, Lu y Bolt (2015) ofrecen un modelo de respuesta al ítem multidimensional a dos niveles aportando un modo informativo de estudiar los efectos (o su ausencia) de la variabilidad entre países en estilos de respuesta.

El reto verdadero tal como Kubinger, Hohensinn, Hofer et al. (2011) exponen es hacer frente a restricciones impuestas por determinadas variables moderadoras tales como formatos de respuestas diferentes y tópicos de contenido diversos. PISA administra el mismo ítem en diferentes posiciones dentro del cuadernillo/libreto; por tanto, la ocurrencia de efectos de posición influirá crucialmente en la dificultad del ítem. No tener en cuenta los efectos del aprendizaje en la toma del test, de la motivación para responder y de la fatiga del examinando originará resultados con un sesgo en la estimación del índice de dificultad del ítem (Hohensinn, Kubinger, Reif, Schleicher & Khorramdel, 2011).

La naturaleza plurilingüística de los tests

La naturaleza plurilingüística de los tests, que deben traducirse a múltiples lenguas, hace cuestionable que sean formas no ya equivalentes sino al menos paralelas de un mismo instrumento ante la diversidad de contextos culturales. De hecho, sobre un total de 101 versiones nacionales de los materiales de Lectura que se usaron en PISA 2009 para 45 idiomas, sólo se realizó un examen profundo de los instrumentos con las versiones en francés e inglés, reconociéndose la especial complejidad de las traducciones a lenguas no indo-europeas como el chino y el árabe (OECD, 2012, p. 813). Grisay, de Jong, Gebhardt, Berezner y Halleux-Monseu (2007) ponen de manifiesto que, con PISA y en general con los estudios comparados internacionales, nadie debiera sorprenderse de que los instrumentos cognitivos son más apropiados en términos culturales y lingüísticos para los países occidentales, que constituyen la mayoría de los países participantes.

PISA 2012 (OECD, 2014b; Cap. V) mantiene la doble traducción y validación métrica siguiente en solo dos idiomas fuente (inglés predominantemente, y francés). Sin embargo, versiones nacionales están disponibles en español buscando una extraña adaptación idiosincrática; quizá la lengua española está floreciendo, como le acaeció al

latín en el pasado, en múltiples nuevos idiomas.

Confusión entre validez y fiabilidad

PISA 2009, en concreto para el caso español (ver Ministerio de Educación, 2010) e inglés (Jerrim, 2011), ofrece una muy cuestionable validez de los instrumentos a través de un proceso de validez relacionada con el criterio por concurrencia y calculada mediante "la correlación entre los promedios de países y comunidades autónomas en PISA 2009, y entre estos resultados y los de 2006" (p. 18); alcanzándose ufanamente índices superiores a 0.86.

Se ignora que la validez concurrente relativa a criterio exige una variable externa pertinente, no contaminada de una administración a otra y ya aceptada como válida, y no los valores de anclaje de ítems anteriores, cuando además estos valores de anclaje son escasos e inadecuados. PISA 2009 en verdad lo que realiza así es una determinación de la fiabilidad por estabilidad, también deseable para ganar validez, mediante un proceso de test-retest (i. e. edición 2006 versus 2009) sobre la base de ítems comunes de enlace entre dos aplicaciones separadas en el tiempo. La determinación de la fiabilidad "oficial" de los instrumentos se realiza en definitiva por consistencia interna de unidades (ítems) variando por dominios (Lectura, Matemáticas y Ciencias, tanto en papel como digital) y según el método, si puntuaciones de máxima verosimilitud (entre 0.75 y 0.85) o valores plausibles (entre 0.30 y 0.86) (OECD, 2012, p. 194). La plétora de coeficientes de fiabilidad que PISA suele ofrecer no es garantía de validez de los instrumentos (OECD, 2012, pp. 234-238).

No correspondencia entre contenidos de los instrumentos y currículos nacionales

El contenido de la pruebas PISA se define antes bien por el juicio de expertos nacionales (un supervisor por país) que por una derivación de los contenidos curriculares nacionales; aunque PISA se sustente sobre un marco curricular (una teoría del currículo) funcional no siempre coincidente plenamente con los currículos convencionales, centrados en el logro de conocimientos más académicos. La consecuencia de tan emergente y sutil currículo para las políticas educativas nacionales es clara: adaptemos los currículos a las demandas de PISA; la evidencia de cambios curriculares auspiciados por PISA es abundante: Chisholm (2015) en Alemania y Sudáfrica; Dolin y Krogh (2010) in el contexto danés, y especialmente en Turquía (Gur, Celik & Ozoglu, 2012) donde las autoridades educativas ya decidieron realizar reformas del viejo currículo mucho antes de antes de que saliesen los resultados de PISA 2003, y que después utilizaron para justificar tales reformas. Debe entonces comprenderse que, aunque PISA es una encuesta intencional a gran escala, es difícil pero no imposible y quizás necesario conectarla con los currículos nacionales.

Añádasele la presencia de sesgos en ítems, detectada por Olsen y Lie (2011) y Kjaernsli y Lie (2011), debida a diferencias culturales, de género, de la dualidad lengua materna y de enseñanza o de tradiciones de enseñanza, que se ignoran, o que PISA opera con constructos latentes generados por ítems validados en muestras bastante reducidas

No garantía de administración estandarizada

Una cuestión general asociada al *testing* y como tal relacionada con PISA es el uso de preguntas-ítems abiertas lo cual suscita problemas de baja fiabilidad, necesidad de concordancia en los correctores/codificadores por dificultadas en la corrección de respuestas muy variadas y tiempo discrecional asignado a la tarea en el acto del examen.

No obstante, PISA también suele utilizar ítems cerrados de elección múltiple (con cuatro alternativas, la mayoría, o cinco alternativas) sobre los que no tenemos constancia de descuento por azar en las puntuaciones de acierto; o sea, no se informa si se aplica la archiconocida corrección de Lafourcade (1971) a la puntuación final aditiva de tales ítems.

No hay garantías en la estandarización de la administración de las pruebas pues cierta opacidad persiste en los informes, que no especifican si los administradores y correctores/codificadores de los tests son externos o internos a los centros, aunque los materiales se remitían a la autoridad del centro (director) que no impartía docencia a alumnos de 15 años. Este matiz pudiera ser crucial, y aunque PISA (OECD, 2012, p. 24) pretendidamente garantiza la estandarización en la administración, captación y procesamiento de datos, surgen dudas de que los datos generados sean comparables entre países dada la diversidad y multiplicidad de ámbitos ya que PISA "da la oportunidad a los países participantes de adaptar ciertos procedimientos y cuestiones para ajustarlos a las circunstancias locales y añadir componentes opcionales que son únicos para un contexto nacional particular" (OECD, 2012, p. 148), de aquí que las mayoría de países omitieron ciertos ítems y administraron otros determinados ítems y libretos (OECD, 2012, Tabla 12.8; pp. 195-187).

El uso reciente, pero previsiblemente generalizado para PISA 2015, de administraciones computerizadas sólo introduce confusión ya que no se puede dilucidar si el desempeño responde a la capacidad del alumno en un área de contenido o en la falta de acceso y manejo del ordenador; algo de esto podría haber sucedido con los resultados de España en PISA 2012 para Resolución de problemas (OECD, 2014, p. 1).

El constructo funcionalista establecido en los ítems de PISA sería también bastante discutible pues tales ítems son en definitiva estereotipados, y más aún los de elección múltiple que también se proponen, como cualquier prueba evaluativa estandarizada y más próximos a los de un test de inteligencia, con una fuerte carga del factor verbal, que a un tarea de una materia escolar.

Artefactos en la estimación de puntuaciones agregadas

PISA no permite obtener puntuaciones individuales consistentes de la unidad básica de análisis: el alumno de 15 años, dado el reducido tamaño de las pruebas de evaluación que se le administran a éste; de ahí su inutilidad como prueba diagnóstico del alumno. La objeción sería clara: si las puntuaciones agregadas no sirven, difícilmente lo serían las poblacionales; o sea, el salto, sujeto→muestra→ población es harto cuestionable por mucho que se apele y ejecuten procedimientos estadísticos sofisticados sobre ítems necesariamente discriminativos sin verificar el supuesto de normalidad multivariada.

La mera agregación de puntuaciones en cada ítem, para los que sí existe criterio de acierto y error, conduce a la puntuación agregada sobre la que no se establece criterio de valía alguno; en última instancia, se opta por la evaluación relativa a norma frente a una más deseable de dominio-criterio.

Uso ventajista de puntuaciones transformadas por tipificación

PISA aporta básicamente puntuaciones transformadas mediante una escala tipificada/estandarizada de transformación del tipo $x_t = 500 + 100\ x_i$; donde x_t es la puntuación típica transformada CEEB (*College Entrance Examination Board*); pero no suministra las directas, ni las totales según grupos ni desviación típica asociada alguna. La potencialidad de las puntuaciones CEEB radica en que permiten, como toda puntuación estandarizada, realizar comparaciones al margen del tamaño diverso de los instrumentos de medida.

Pero al generar puntuaciones transformadas se magnifican las diferencias y por ende una falsa varianza entre puntuaciones. Una de las consecuencias de tal transformación es denotar diferencias supuestamente amplias que en realidad no existen, pues éstas son puro artefacto métrico. Se cae en una de las limitaciones del operacionalismo, que degenera conforme avanza en sus operaciones progresivas cual sistema de alta entropía, que van perdiendo su sentido: realizar una evaluación del dominio competencial.

El intento posterior de una adscripción en PISA 2009 Lectura a ocho niveles para-criteriales de menor a mayor rendimiento (*<1b*, *1b*, *1a*, *2*, *3*,*4*,*5* y *6*) o los usuales siete niveles en Matemáticas y Ciencias (*<1*, *1*, *2*, *3*, *4*, *5*, *6*), y más aún los *rankings* de naciones o comunidades, que con estos niveles se realizan, sigue sin resolver justificadamente en base a evidencia ulterior qué valor de la puntuación agregada o nivel serían deseable para definir el nivel competencial que ha alcanzado un alumno o subgrupo poblacional determinados; no obstante, se asuma liberalmente que: "los estudiantes que se encuentran en el nivel de rendimiento *2* en Lectura demuestran el tipo de competencia mínimo requerido para el aprendizaje posterior y la vida social y laboral (MECD, 2012, p. 61). Sea el nivel consensuado *2* o el más deseable y contundente nivel *3*, es difícil ubicar descriptores de competencia en el nivel apropiado.

En el fondo, PISA pretende combinar subrepticiamente dos tipos de evaluación: la relativa a norma dada por el desempeño grupal con una pretendida relativa a criterio de expertos, que ordinaliza la calidad de los proceso cognitivos modelados previamente como de alta o baja demanda cognitiva y jerarquizados después en niveles de desempeño (1 a 6) pero que en verdad tales niveles se generan por valores relativos a norma. Así, para España, el intervalo de puntuaciones transformadas asociadas a cada nivel en PISA 2009 Lectura es: 1b [262-334]; 1a [335-406]; 2 [407-479]; 3 [480-552]; 4 [553-625]; 5 [626-707]; 6 [≥708].

Confianza reverencial en la significación estadística como técnica de análisis de datos

PISA ha venido incorporando la significación, solamente estadística que no sustantiva, de las diferencias inter-países, intra-países, entre otros subgrupos poblacionales y variables moderadoras consideradas, a partir de componentes de varianza en la más pura línea fisheriana (Fisher, 1925). Igualmente, las diferencias entre sucesivas ediciones-trienios siguen confiando en el "sacrosanto nivel de salvación" de $\alpha = 0.05$ y omitiendo a cualquier consideración a la potencia del contraste y a los tamaños del efecto; no hay constancia alguna de "cuánta es la diferencia que marca la diferencia" (Fernández-Cano & Fernández-Guerrero, 2009; pp. 101-102). Recuérdese que obtener diferencias estadísticamente significativas es bien sencillo pues basta con aumentar el tamaño muestral, ya alto en estudios de encuesta a gran escala como PISA.

Asombra que informes tan sesudos de tan caros programas evaluativos sigan lastrados en la simplicidad e irrelevancia de mantener el uso de la diferencia bruta en las puntuaciones medias, sólo porque tal diferencia sea estadísticamente significativa según un contraste inferencial o correlacional usando un estadístico de significación, que no siempre se declara (si *z*, *F*, *t* o χ^2).

Toda la significación de los resultados es totalmente estadística y multipresente en los diversos informes. Al identificar la significación estadística con la significación sustantiva se vicia la investigación básica que generó este programa, la investigación aplicada que los desarrolla y los estudios derivados por PISA, los cuales siguen reverencialmente confiando en la significación estadística.

Ausencia de estadísticos de significación sustantiva como magnitudes del efecto

Debe recordarse que desde su cuarta edición, el Manual de Publicación de la *American Psychological Association* (1990) enfatiza que los valores *p* no son índices aceptables del efecto y "anima" a informar con tamaños del efecto.

Las múltiples comparaciones intergrupos expuestas en PISA deberían utilizar al menos un estadístico del tamaño del efecto paramétrico calculado a partir de una diferencia entre medias estandarizada (i. e. la ubicua *d* de Glass, 1977) u otro estadístico de naturaleza correlacional como el coeficiente de correlación intraclase R_I que determine el porcentaje de varianza explicada; para conceptualizaciones y fórmulas sobre

magnitud del efecto, véase Fernández-Cano y Fernández-Guerrero (2009, pp. 99-124).

Ante el viejo truco de que para alcanzar significación estadística abasta aumenta el tamaño de los grupos a comparar, necesitamos entonces valores criteriales para denotar cuánta es la diferencia que marca la diferencia, determinar el tamaño muestral conveniente considerando al par nivel de significación (α) y potencia del contraste (1-β). Un extraño valor de diferencia de 9 puntos de la escala CEEB (500; 100) podría estar considerándose significativo aunque no hay evidencias ni autor que la proponga a no ser la Wikipedia (2014): ".. una diferencia de 9 puntos es suficiente para ser considerada significativa". [sic]

No obstante, a partir de los datos de PISA se pueden calcular tamaños del efecto. Lynn y Mikk (2009) con datos de PISA 2000, 2003 y 2006 y datos de PIRLS 2001 y 2006 calculan un tamaño del efecto promedio (d = 0.42) a favor de las alumnas en Lectura; valor casi similar al 0.44 obtenido con datos PISA 2009 por Reilly (2012), mientras que en Matemáticas los alumnos se desempeñan mejor que las alumnas con un d = 0.22.

Problemática presentación de hallazgos

Los informes PISA suelen ser excesivamente técnicos, con un lenguaje burocrático, que abundan en tablas y grafos de enorme extensión y prolijidad; por ejemplo, en los informes técnicos de 2012 (OECD, 2014b) se exponen 282 tablas y 75 figuras. *PISA* abunda en el uso tal vez excesivo de *rankings* y estadísticos diversos poco inteligibles incluso para el lector cualificado del campo de la educación y dejando que tales hallazgos sean proclives a malas interpretaciones.

En esta línea, Berliner (2015) comenta que tratar de comprender PISA es análogo a la parábola de los ciegos y del elefante en la que la más importante de todas las cuestiones asociadas se discute, en concreto la interpretación de la puntuaciones de la naciones, sin llegar a una consideración global. Takayama (2008) habla de descontextualizadas "tablas de ligas internacionales" [entiéndase al modo deportivo]. Lee (2014) clarifica que los rankings PISA públicamente emitidos cambian en cierto grado si son reevaluados considerando otros factores.

Cuestionables implicaciones de los hallazgos para la normativa y la práctica educativas

Sobre los hallazgos de *PISA* se ha pretendido establecer implicaciones que orienten la normativa y las políticas educativas. Sin embargo, se hace difícil establecer inferencias evaluativas consistentes sobre todo un sistema escolar utilizando datos transversales y limitados al nivel de los alumnos de 15 años pretendidamente muestreados por *PISA*. Su *alter ego* (NAEP) trabaja con tres cohortes: alumnos de 4º, 8º y 12º grado; un marco poblacional más amplio y diversificado.

Atención debiera prestarse además al perverso sesgo de agregación (falacia ecológica) en las interpretaciones que rotula a todos los sujetos, incluso a aquellos capaces, por estar incluidos en el grupo o país más desaventajado y en especial a la consideración del actual pánico que rodea el bajo rendimiento de chicos (Smith, 2009). Igualmente, las comparaciones entre sistemas nacionales de educación son forzadas.

Pese a voceríos desproporcionados por magnificados en los medios de comunicación y círculos políticos, algunos autores (i. e. Yore, Anderson y Chiu, 2010) consideran que ha sido escaso el impacto propiamente educativo de *PISA* sobre el desarrollo del currículo, la enseñanza del profesor, la evaluación específica de alumnos, profesores y centros y las políticas educativas, quedando tan breve transferencia en un mero análisis de normativas (*policy analysys*)[6].

Conclusiones y recomendaciones

PISA constituye un indudable esfuerzo evaluativo que ha generado abundante investigación, eminentemente cuantitativa, aunque se echan en falta más estudios cualitativos, y los disponibles suelen estar

insertos en el paradigma socio-crítico; pero PISA necesita ejercer un mayor rigor metodológico y declarar claramente sus métodos en futuros informes, sobre todo en los informes nacionales emitidos periódicamente. PISA necesita prestar una atención creciente al desarrollo de datos longitudinales, algo que ya se reconocía como objetivo en PISA 2012. Para ello, antes que elaborar comparaciones entre dos aplicaciones (OECD, 2013, p. 8) debiera elaborar tendencias longitudinales como series temporales para mostrar la evolución de los sistemas educativos en un plano comparativo internacional; pero habría que tener la reserva de que los instrumentos de una edición a otras sean iguales o paralelos, pues se han ido descartando algunos, introduciendo nuevos ítems y modificaciones en los anteriores.

PISA es sólo un macro estudio evaluativo internacional; ante ello, habría entonces que ser muy cuidadoso en establecer inferencias infundadas, por pretendidamente derivadas de él, relativas a su nula función diagnóstica del desempeño individual del alumno, a cambios curriculares, a comparaciones reduccionistas no contextualizadas o a prácticas del profesor. No obstante, el impacto de PISA es fuerte y alentador con notables consecuencias positivas; por consiguiente, su estudio merecería una indagación más profunda y continuada de carácter meta-evaluativo. Un modo de contribuir a la mejora de un programa es oír la críticas que se le hacen, la cuales no siempre son fáciles de formular.

Como corolario evaluativo final bien pudiera decirse: SÍ a PISA, *ma non troppo*.

Referencias

Adams, R. J. (2003). Response to 'Cautions on OECD's recent educational survey (PISA)'. *Oxford Review of Education, 29*(3), 377-389. doi: http://dx.doi.org/10.1080/0305498032000120319

Adams, R., Berezner, A. & Jakubowski, M. (2010). *Analysis of PISA 2006 preferred items ranking using the percent correct method*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2006/44919855.pdf

Adams, R. J., Wu, M. L. & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessments. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 271–280). New York: Springer.

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: APA.

Bank, V. (2012). On OECD policies and the pitfalls in economy-driven education: The case of Germany. *Journal of Curriculum Studies*, *44*(2), 193-210. doi: http://dx.doi.org/10.1080/00220272.2011.639903

Berliner, D. C. (2011). The context for interpreting PISA results in the USA. Negativism, chauvinism, misunderstanding, and the potential to distort the educational systems of nations. In M. Pereira, H-G. Kotthoff & R. Cowen (Eds.), *PISA under examination: Changing knowledge, changing tests, and changing schools* (pp. 77-96). Rotterdam: Sense Publishers.

Berliner, D. C. (2015). The many facets of PISA. *Teachers College Record*, *117*(1), 20.

Bollen, K., Paxton., P. & Morishima, R. (2005). Assessing international evaluations. An example from USAID's democracy and governance program. *American Journal of Evaluation*, *26*(2), 189-203. doi: http://dx.doi.org/10.1177/1098214005275640

Brown, G., Micklewright, J., Schnepf, S. V. & Waldmann, R. (2007). International surveys of educational achievement: How robust are the findings? *Journal of the Royal Statistical Society Series A-Statistics in Society*, *170*(3), 623-646. doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00439.x

Carnoy, M. & Rothstein, R. (2013). *International tests show achievement gaps in all countries, with big gains for U.S.*

disadvantaged students. Economic Policy Institute: Washington, DC. Retrieved from http://www.epi.org/blog/international-tests-achievement-gaps-gains-american-students/

Couso, D. (2009). Y después de PISA, ¿qué? Propuestas para desarrollar la competencia científica en el aula de ciencias [And after PISA, what? Proposals to develop the scientific competence in the Science classroom]. *Enseñanza de las Ciencias*, especial issue, 3547-3550.

DeSeCo Project (2008). *Definition and selection of competencies: Theoretical and conceptual foundations*. Paris: OECD. Retrieved from http://www.deseco.admin.ch/

De Witte, K. & Kortelainen, M. (2013). What explains the performance of students in a heterogeneous environment? Conditional efficiency estimation with continuous and discrete environmental variables. *Applied Economics*, *45*(17), 2401-2412. doi: http://dx.doi.org/10.1080/00036846.2012.665602

Dolin, J. & Krogh, L. B. (2010). The relevance and consequences of PISA science in a Danish context. *International Journal of Science and Mathematics Education*, *8*(3), 565-592. doi: http://dx.doi.org/10.1007/s10763-010-9207-6

Drechsel, B., Carstensen, C. & Prenzel, M. (2011). The role of content and context in PISA interest scales: A study of the embedded interest items in the PISA 2006 science assessment. *International Journal of Science Education*, *33*(1), 73-95. doi: http://dx.doi.org/10.1080/09500693.2011.518646

Ehmke, T., Drechsel, B. & Carstensen, C. H. (2008). Klassenwiederholen in PISA-I-Plus: Was lernen sitzenbleiber in mathematik. [Grade repetition in PISA-I-Plus: What do students who repeat a class learn in mathematics?]. *Zeitschrift fur Erziehungswissenschaft*, *11*(3), 368-387. doi: http://dx.doi.org/10.1007/s11618-008-0033

Ercikan, K., Roth, W-M. & Asil, M. (2015). Cautions about inferences from international assessments: The case of PISA 2009. *Teacher College Records*, *117*(1), 1-28.

Fernández-Cano, A. & Fernández-Guerrero, IM. (2009). *Crítica y alternativas a la significación estadística en el contraste de hipótesis*. Colección Cuadernos de Estadística, nº 37. Madrid: Arco Libros-La Muralla.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*(1), 351-379.

Grisay, A., De Jong., Gebhardt, E., Berezner, A. & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249-266

Gur, B. S., Celik, Z. & Ozoglu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, *27*(1), 1-21. doi: http://dx.doi.org/10.1080/02680939.2011.595509

Hanberger, A. (2014). What PISA intends to and can possibly achieve: A critical programme theory analysis. *European Educational Research Journal*, *13*(2), 167-180. doi: http://dx.doi.org/10.2304/eerj.2014.13.2.167

Hartig, J. & Frey, A. (2012). Validity of a standard-based test for mathematical competencies. Relations with the competencies assessed in PISA and variance between schools and school tracks. *Diagnostica*, *58*(1), 3-14. doi: http://dx.doi.org/10.1026/0012-1924/a000064

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E. & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, *17*(6), 497-509. doi: http://dx.doi.org/10.1080/13803611.2011.632668

IEA (2012). *The International Association for the Evaluation of Educational Achievement*. Retrieved from http://www.iea.nl/

Jerrim, J. (2011). *England's" plummeting" PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline* (Nº. 11-09). London: Department of Quantitative Social Science-Institute of Education of University of London.

Judkins, D. R. (1990). Fay's method of variance estimation. *Journal of Official Statistics*, *6*, 223-239.

Kjaernsli, M. & Lie, S. (2011). Students' preference for science careers: International comparisons based on PISA 2006. *International Journal of Science Education*, *33*(1), 121-144. doi: http://dx.doi.org/10.1080/09500693.2011.518642

Knipprath, H. (2010). What PISA tells us about the quality and inequality of Japanese Education in Mathematics and Science. *International Journal of Science and Mathematics Education*, *8*(3), 389-408.

Kreiner, S. & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210-231. doi: http://dx.doi.org/10.1007/s11336-013-9347-z

Kubinger, K. D., Hohensinn, C., Hofer, S, Khorramdel, L, Freborta, M, Holocher-Ertl, S., Reif, M. & Sonnleitner, P. (2011). Designing the test booklets for Rasch model calibration in a large-scale assessment with reference to numerous moderator variables and several ability dimensions. *Educational Research and Evaluation*, *17*(6), 483-495. doi: http://dx.doi.org/10.1080/13803611.2011.632666

Lafourcade, P. (1971). *Evaluación de los aprendizajes* [Learning evaluation]. Buenos Aires: Kapelusz.

Lee, J. (2014). An attempt to reinterpret student learning outcomes: A cross-national comparative study. *Peabody Journal of Education*, *89*(1), 106-122. doi: http://dx.doi.org/10.1080/0161956X.2014.862476

Lu, Y. & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-scale Assessments in Education*, *3*(2). doi: http://dx.doi.org/10.1186/s40536-015-0012-0

Lynn, R. & Mikk, J. (2009). Sex differences in reading achievement. *Trames-Journal of the Humanities and Social Sciences, 13*(1), 3-13. doi: http://dx.doi.org/10.3176/tr.2009.1.01

Meyer, H-D. & Benavot, A. (Eds.). (2013). *PISA, power, and policy. The emergence of global educational governance*. Providence, RI: Symposium Books.

Ministerio de Educación, Cultura y Deporte [MECD] (2010). *PISA 2009. Programa para la Evaluación Internacional de los Alumnos. OCDE. Informe español* [PISA 2009.The Spanish report]. Madrid: Instituto de Evaluación. Retrieved from http://www.educacion.gob.es/dctm/ministerio/horizontales/prensa/notas/2010/20101207-pisa2009-informe-espanol.pdf?documentId=0901e72b806ea35a

National Center for Education Statistics (2012). *National Assessment of Educational Progress*. Retrieved from http://nces.ed.gov/nationsreportcard/

Olsen, RV. & Lie, S. (2011). Profiles of students' interest in science issues around the world: Analysis of data from PISA 2006. *International Journal of Science Education*, *33*(1), 97-120. doi: http://dx.doi.org/10.1080/09500693.2011.518638

Organisation for Economic Co-operation and Development (2009). *PISA 2009 key findings*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2009/pisa2009keyfindings.htm

Organisation for Economic Co-operation and Development (2012). *PISA 2009 Technical report*, PISA. Paris: OECD Publishing. doi: http://dx.doi.org/10.1787/9789264167872-en

Organisation for Economic Co-operation and Development (2013a). *PISA 2012 Results in focus. What 15-year-olds know and what they can do with what they know*. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

Organisation for Economic Co-operation and Development (2013b). *PISA 2012 results*. What make schools successful? Resources, policies and practices. Vol. 4. Paris: OECD. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-IV.pdf

Organisation for Economic Co-operation and Development (2014a). *SPAIN – Country note –Results from* PISA *2012 problem solving*. Retrieved from http://www.oecd.org/spain/PISA-2012-PS-results-eng-SPAIN.pdf

Organisation for Economic Co-operation and Development (2014b). *PISA 2012 technical report*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Organisation for Economic Co-operation and Development (2015a). *School governance, assessments and accountability*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/keyfindings/Vol4Ch4.pdf

Organisation for Economic Co-operation and Development (2015b). *PISA 2012 results*. Paris: OECD. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm

Prais, S. J. (2003). Cautions on OECD'S recent educational survey (PISA). *Oxford Review of Education*, *29*(2), 139-163. doi: http://dx.doi.org/10.1080/0305498032000800657

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *Plos One*, *7*(7). doi: http://dx.doi.org/10.1371/journal.pone.0039904

Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education*, *27*(2), 115-132. doi: http://dx.doi.org/10.1080/08957347.2014.880440

Rychen, D. S. & Salganik, L. H. (2003). *Key competencies for successful life and a well-functioning society*. Göttinga: Hogrefe & Huber.

Scriven, M. (2011). *Evaluating evaluations:A meta/evaluation checklist*. (6th ed.). Retrieved from http://michaelscriven.info/images/EvaluatingEvals-Checklist.pdf

Smith, E. (2009). Underachievement, failing youth and moral panics. *Evaluation & Research in Education*, *23*(1), 37-49.

Strietholt, R., Rosén, M. & Bos, W. (2013). A correction model for differences in the sample compositions: the degree of comparability as a function of age and schooling. *Large-scale Assessments in Education*, *1*(1). doi: http://dx.doi.org/10.1186/2196-0739-1-1

Stufflebeam, D. (2011). Meta-evaluation. *Journal of MultiDisciplinary Evaluation*, *7*(15), 99-158.

Takayama, K. (2008). The politics of international league Tables: PISA in Japan's achievement crisis debate. *Comparative Education*, *44*(4), 387-407. doi: http://dx.doi.org/10.1080/03050060802481413

Wang, M.C., Haertel, G.D. & Walberg, H.J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63*(3), 249-294. doi: http://dx.doi.org/10.3102/00346543063003249

Wikipedia (2014). *Informe* PISA [PISA Report]. Retrieved from http://es.wikipedia.org/wiki/Informe_PISA

Yarbrough, D.B., Shulha, L.M., Hopson, R.K. & Caruthers, F.A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Yore, L. D., Anderson, J. O. & Chiu, M. H. (2010). Moving PISA results into the policy arena: Perspectives on knowledge transfer for future considerations and preparations. *International Journal of Science and Mathematics Education*, *8*(3), 593-609.

Zuckerman, G. A., Kovaleva, G. S. & Kuznetsova, M. I. (2013). Between PIRLS and PISA: The advancement of reading literacy in a 10-15-year-old cohort. *Learning and Individual Differences*, *26*, 64-73. doi: http://dx.doi.org/10.1016/j.lindif.2013.05.001

Notas

1. Especialmente en países latino-americanos (México, Argentina, Brasil y Uruguay), en Egipto o Tailandia la escolaridad obligatoria termina a los 14 años. Otros países donde el término de la escolaridad se pospone a los 15 años o incluso a los 16 años, cuentan con tasas bastante altas de trabajo infantil y abandono escolar.
2. Las limitaciones de las técnicas de remuestreo son abundantes. Tres de ellas se exponen a continuación. En primer lugar, representan un canje pobre y potencialmente desconcertante de la cosa real pues, aunque estos análisis estén basados en submuestras independientes extraídas de la muestra original, siempre están restringidos a las características de la muestra original; en consecuencia no son el sustituto de una replicación verdadera. No tienen en cuenta las potenciales desviaciones y sesgos ocultos asociados a todo estudio transversal, basado en una única muestra de participantes que genera una población sobre la que se asumen ciegamente ciertos supuestos. Además, determinan evaluaciones inflacionadas por operar con muestras poco representativas y muy dependientes.
3. Personalmente hubiese optado por bootstrap, ya que al considerar tantas configuraciones de sujetos, en las que un mismo caso podría estar representado varias veces o ninguna, se puede denotar hasta qué punto los resultados son estables y generalizables a través de diferentes tipos de sujetos. Además, no es necesario que en la muestra generadora se verifiquen los supuestos de parametricidad, manteniendo la medida interval, y/o cuando no podemos asumir modelo alguno sobre la distribución poblacional, pues es factible realizar bootstrap no paramétrico.
4. No parece apropiado considerar a NAEP un "equivalente" a PISA ya que ambos no están relacionados de algún modo y sirven a fines diferentes. No obstante, PISA podría aprender de NAEP su diseño longitudinal y su énfasis curricular.
5. Un caso inverso es el expuesto por Hartig y Frei (2012) en el que los datos de PISA 2006 son la variable criterio para denotar la validez concurrente curricular del test basado en estándares de competencias matemáticas utilizado en el sistema educativo alemán. Tal test correlaciona con PISA Matemáticas (r = 0.94), con PISA Lectura (r = 0.75) y PISA Ciencias (r = 0.85). Obviamente, los autores reconocen la imagen incompleta de la validez de tal test en base a sólo las correlaciones extraídas.
6. PISA se cura en salud (OECD, 2012, p. 3) al manifestar que: "PISA examina cómo de bien los estudiantes están preparados para hacer frente a los retos del futuro, antes que cómo de bien ellos dominan currículos particulares" (p. 3) y "mira la capacidad de estos para usar su conocimiento y destrezas ante retos de la vida real" (p. 22). PISA induce un alumno "standard" considerando que la evaluaciones completas a nivel de

alumno son innecesarias y sumergiendo su componente dentro de la evaluación nacional sin generar puntuaciones para los individuos, incluso ni lo pretenden; por ello la falacia ecológica está siempre presente en los múltiples informes al permitir la confusión del grupo (estudiantes) con el caso (un/a estudiante).

Author / Autor **To know more / Saber más**

Fernandez-Cano, Antonio (afcano@ugr.es).

Catedrático en el Departamento de Métodos de Investigación y Diagnóstico en Educación de la Facultad de Ciencias de la Educación. Universidad de Granada (España). Sus principales áreas de interés son las metodologías de investigación y evaluación, cienciometría e investigación educativa. Su dirección postal es: Universidad de Granada. Facultad Ciencias de la Educación. Departamento de Métodos de Investigación y Diagnóstico Educativo. Despacho 216.1 Campus de Cartuja. 18071- Granada (España)

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]