Taut, Sandy & Palacios, Diego (2016). Interpretaciones no intencionadas e intencionadas y usos de los resultados de PISA: Una perspectiva de validez consecuencial. *RELIEVE*, *22*(1), art. M8. DOI: http://dx.doi.org/10.7203/relieve22.1.8294

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Interpretaciones no intencionadas e intencionadas y usos de los resultados de PISA: Una perspectiva de validez consecuencial

Entended and unintended interpretations and uses of PISA results: A consequential validity perspective

Taut, Santy [(1)]; Palacios, Diego [(2)]

(1) Pontificia Universidad Católica de Chile (2) Pontificia Universidad Católica de Chile

Resumen

Este artículo explora la relevancia de considerar las consecuencias de las pruebas como parte de las discusiones acerca de la validez, la investigación sobre validación, en el contexto del Programa Internacional para la Evaluación de Estudiantes de la OCDE, PISA. Lo primero que describe la concepción moderna de validez como un aspecto fundamental de la calidad de los ensayos y sistemas de pruebas, es que evoluciona en torno a las interpretaciones propuestas y usos de las puntuaciones de las pruebas: "La validez se refiere al grado en el cual la evidencia y la teoría apoyan las interpretaciones de las puntuaciones de las pruebas sobre las propuestas de su uso en los test. La validez es, por tanto, la consideración más fundamental en el desarrollo y evaluación de las pruebas". (AERA, APA & SNEM, 2014, p. 11). En particular, nos centramos en el papel que han jugado sus consecuencias en la literatura sobre validez de la prueba y validación. Así como a continuación, introducimos PISA y sus interpretaciones y usos previstos como base para el examen de su validez. Esto es seguido por un resumen de los estudios empíricos existentes sobre los usos y consecuencias de PISA. Finalmente, el documento presenta piezas que faltan en la evidencia de validez en relación con las consecuencias y se analiza la importancia de una agenda pro-activa en estos temas por parte de los grupos de interés de PISA a nivel internacional y nacional

Fecha de recepción
1 Abril 2016

Fecha de aprobación
22 Junio 2016

Fecha de publicación
22 Junio 2016

Palabras clave:

PISA; validez; usos puntuaciones de test; validación; construcción de pruebas

Abstract

This paper explores the relevance of considering the consequences of testing as part of discussions about the validity, and validation research, in the context of the OECD Programme for International Student Asssessment, PISA. We first describe the modern conception of validity as a core aspect of quality of tests and testing systems, evolving around the proposed interpretations and uses of test scores: “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests.” (AERA, APA & NCME, 2014, p. 11). In particular, we focus on the role that consequences have played in the literature on test validity and validation. We then introduce PISA and its intended interpretations and uses as the basis for examining its validity. This is followed by summarizing existing empirical studies on the uses and consequences of PISA. Finally, the paper presents missing pieces in the validity evidence related to consequences and discusses the importance of a pro-active agenda on these topics by the PISA stakeholders at international and national levels.

Reception Date
2016 April 1

Approval Date
2016 June 22

Publication Date:
2016 June 22

Keywords:

PISA; validity; uses of test scores; validation; developing tests.

Autor de contacto / Corresponding author

Taut, Sandy. Pontificia Universidad Católica de Chile, Escuela de Psicología, Centro de Medición MIDE UC, Avda Vicuña Mackenna 4860, Macul, Santiago (Chile). staut@uc.cl

Validez consecuencial: Introducción

Los Estándares educativos y pruebas psicológicas (AERA, APA y NCME, 2014), denominados en lo sucesivo "los estándares", representan un consenso profesional sobre los criterios para juzgar la calidad de la educación y las pruebas psicológicas en los Estados Unidos. A pesar de que fueron desarrollados en los Estados Unidos, se han convertido en una influyente referencia utilizada por las comunidades de medición fuera de los EE.UU. Los estándares definen tres aspectos centrales de pruebas de calidad: validez, confiabilidad/precisión y errores de medición, y equidad. Entre estos tres, validez ocupa el centro del escenario y será el foco de este artículo. Concretamente, el documento se centra en un tipo particular de evidencia de validez, los estándares que se recomiendan incluir en las pruebas de validación: evidencia basada en las consecuencias de las pruebas.

Entre otras fuentes de evidencia de validez, los Estándares mencionan las consecuencias y distinguen entre las consecuencias que se derivan directamente de las "interpretaciones y usos de las puntuaciones de las pruebas previstas por desarrolladores de pruebas", "alegaciones sobre el uso de pruebas que no están directamente sobre la base de interpretaciones de puntuación de prueba" y "consecuencias que son involuntarias" (2014, pp. 19-20). Esas interpretaciones y usos previstos por el desarrollador de la prueba deben ser validados por el desarrollador de la prueba, proporcionando evidencia teórica y/o empírica relevante para cada uno. Sin embargo, los Estándares señalan que "evidencias sobre las consecuencias es relevante para validar cuando puede atribuirse a una fuente de invalidez como construir una representación insuficiente o construir componentes irrelevantes. La evidencia de que no se puede ser rastreado no es relevante para la validación de la interpretación de las puntuaciones de las pruebas" (p. 21) (véase también el Estándar 1.25, págs. 30-31). Por lo tanto, siguiendo a Messick (1989), la invalidez de la prueba se produce sólo cuando las consecuencias se deben a defectos en la prueba, pero no si son externos a las características de la prueba. Cronbach (1988), sin embargo, aboga por un papel más central de consecuencias en las obligaciones de los desarrolladores de pruebas, si no bajo el paraguas de los estudios de validación, como otro tipo de obligación (social) a fin de evaluar la legitimidad del uso de la prueba. En línea con la orientación de Cronbach (1988), un grupo de expertos internacionales encargados de examinar las pruebas de validación de los sistemas de pruebas de rendimiento de los estudiantes de México ENLACE y Excale derivaron un conjunto de criterios de buena práctica en la validación, algo similar a los Estándares (AERA, APA y NCME, 2014), pero formulado de modo más concreto y en un estilo más operativo (Martínez Rizo et al., 2015). Los criterios relacionados con la validez consecuencial se encuentran en el apéndice A.

En los casos en los que el usuario de una prueba propone interpretaciones y usos que difieren de los admitidos por el desarrollador de la prueba, es ese usuario el que tiene la responsabilidad de presentar las correspondientes pruebas de validez (AERA, APA y NCME, 2014, p. 13). Evidencia de validez sobre las consecuencias debe ser presentada para los supuestos que subyacen a la teoría de la acción para cada uso específico de las puntuaciones de pruebas. Por ejemplo, si se planteó que los resultados de PISA se podrían utilizar para vigilar el impacto de una reforma curricular en un país determinado, entonces debe aportarse la evidencia que PISA es una medida válida del impacto de la reforma curricular en ese país. Sin embargo, esto no es responsabilidad del desarrollador de pruebas sino de los encargados de formular las políticas nacionales que utilizan la prueba con este propósito. No obstante, si se plantea que los resultados de PISA, en relación con otros indicadores de la educación medidos mediante cuestionarios de PISA, podrían utilizarse para describir el grado de equidad educativa alcanzado por un determinado país - que

corresponde a una de las pretensiones expresada por la OCDE, el creador de la prueba PISA - entonces el desarrollador de pruebas debe presentar evidencia de que PISA (como los cuestionarios y las pruebas) ofrece realmente información válida sobre la equidad educativa de los países participantes (ver AERA, APA y NCME, 2014, p. 20).

Michael Kane (2006, 2013) desarrolló un enfoque basado en el argumento para la validación. En este enfoque, la validación corresponde a evaluar un argumento interpretativo que consta de un conjunto de supuestos que deben cumplirse para que interpretaciones y usos previstos sean válidos. Con respecto a las consecuencias, Kane señala que muchos programas de ensayos han ido más allá de la tradicional "Función de control, para usar las pruebas como el motor de la reforma y la rendición de cuentas en educación" (2006, p. 55). Este objetivo hace que sean sujetos de validación que debe ser similar a la evaluación de los programas de intervenciones educativas, que incluye los resultados intencionales y no intencionales. En esta concepción de la validación según Kane, los desarrolladores de las pruebas deben analizar las interpretaciones semánticas de los resultados de las pruebas, tanto como los usos de la prueba que explícita o implícitamente recomiendan (Shepard, 1997), mientras que los usuarios de la prueba deben desempeñar un papel importante en el examen de las consecuencias reales del uso de las mismas en su respectivo contexto, población y procedimiento. Los usuarios de las pruebas deben explicar los supuestos subyacentes que conducen a consecuencias deseadas del uso de la prueba, y estos supuestos deben ser creíbles a través de diferentes grupos de interesados. Así, los usuarios de las pruebas deben evaluar estas consecuencias tanto como otras nuevas intervenciones educativas son evaluadas en términos de su eficacia, coste y beneficio, sopesar las consecuencias positivas y negativas (Kane, 2006). Linn (1998) incluye a la Comunidad de medición -en un rol que él califica como "evaluador de test" - entre aquellos que tienen responsabilidades en este sentido.

Sin embargo, también hay quienes han cuestionado si la validez consecuencial en realidad debería ser discutida absolutamente como parte de validez y validación. Por ejemplo, Popham (1997) argumentó que las consecuencias de las pruebas son importantes y deben ser examinadas por ambos desarrolladoresy usuarios de las pruebas, pero no deben ser consideradas parte de validez. Para evitar la confusión de los usuarios de prueba, estos dos conceptos - validez y consecuencias- debe quedar claramente separados. Asimismo, Mehrens (1997) plantea que las consecuencias de las pruebas están más allá del alcance del término "validez": "Vamos a reservar el término para determinar la precisión de las inferencias sobre (y la comprensión de la característica evaluada, no la eficacia de las acciones tras la evaluación." (p. 18)).

La cuestión de las consecuencias como evidencia de validez se complica aún más por el hecho de que "los diferentes tomadores de decisiones pueden hacer diferentes juicios de valor sobre el impacto de las consecuencias en el uso de pruebas." (Kane, 2006, pág. 21) (Véase también Mehrens, 1997). Esto significa que a menudo no existe acuerdo entre los usuarios de prueba, y no verdad objetiva, con respecto a lo que se consideran usos y consecuencias imprevistas y cuáles pueden ser legítimos usos que van más allá de los usos propuestos por el desarrollador de la prueba.

Hoy, en muchas sociedades, las interpretaciones, usos y consecuencias del control de la instrucción reciben gran atención por parte de los agentes educativos, incluidos los maestros y los padres. Asimismo, el aspectoconsecuente de validez en los escritos académicos y marcos de normas parece ahora aquí para quedarse, así que los desarrolladores de pruebas deben incluir su investigación en sus esfuerzos de validación, al menos tan lejos como se pretendía, interpretaciones y usos. Pero no sólo los desarrolladores de pruebas tienen un papel clave en este sentido: los

investigadores pueden apoyar sus esfuerzos por estudiar los usos y consecuencias de los tests en diferentes contextos políticos. Asimismo, los encargados de formular políticas y otros usuarios de prueba deben desarrollar su propia capacidad de comprender completamente las interpretaciones intencionales y no intencionales y los usos de las puntuaciones de las pruebas, y deben resistir mejor las presiones políticas que rodean a prueba y uso en muchos países, a menudo resultando en usos indebidos. Esto también se aplica a los usos de los resultados de PISA, que han servido a las agendas políticas de las autoridades educativas y las fuerzas de oposición, así como a grupos de intereses particulares, a fin de apoyar una variedad de argumentos en un camino no siempre justificado o apropiada desde un punto de vista o de validez según lo previsto por el desarrollador de la prueba.

Los resultados de PISA ¿Qué pinta tienen?

El Programa para la Evaluación Internacional de Alumnos, PISA, de la OCDE es un instrumento de evaluación comparativa internacional de los estudiantes aplicada por la Organización para la Cooperación y el Desarrollo Económico (OCDE) cada tres años, comenzando en el año 2000. Se evalúan los conocimientos y habilidades de alumnos de 15 años en áreas como matemáticas, lectura, ciencias y solución de problemas, dependiendo del año de aplicación. Aproximadamente dos tercios de la prueba -que tarda dos horas en completarse- contiene preguntas abiertas que solicitan a los niños que apliquen sus conocimientos y habilidades a problemas novedosos, mientras que una tercera parte de la prueba contiene elementos de elección múltiple. La evaluación toma una perspectiva de alfabetización y se centra en la capacidad para utilizar los conocimientos y competencias para responder a los desafíos de la vida real, en lugar del dominio específico de los programas escolares. En 2012, más de 500.000 estudiantes en 65 países y economías completaron la evaluación, con un énfasis particular sobre matemáticas. PISA también incluye cuestionarios para estudiantes y para los directores de las escuelas.

Los resultados de PISA se comunican en una variedad de formatos, siendo sus principales destinatarios, siendo los "investigadores, responsables políticos, educadores, padres y estudiantes" sus principales audiencias (véase OCDE, 2014, p. 5). En primer lugar, la OCDE publica los resultados internacionales PISA en diferentes formatos, incluyendo informes completos, resúmenes para políticos, bases de datos, vídeos y presentaciones. Los principales resultados internacionales se publican en un informe de 50 páginas que resume los resultados y las lecciones políticas basadas en el análisis de datos realizados por la OCDE. Este informe contiene una descripción de las principales conclusiones, las tablas mostrando resultados por países enumerados en orden descendente según sus resultados de PISA, gráficos de análisis descriptivo y correlacional y las interpretaciones políticas relevantes de estos análisis como "Fomentar el rendimiento superior y abordar el bajo rendimiento no necesitan ser mutuamente excluyentes" o "la brecha de género en el rendimiento de los estudiantes puede ser reducida considerablemente, en tanto que los niños y las niñas en todos los países y economías muestran que pueden ser exitosos en las tres materias" (OCDE, 2014, p. 9).

Además, un gran número de informes temáticos analizan temas específicos, tales como los factores relacionados con el bajo rendimiento en los distintos países, los posibles orígenes de las diferencias de género, las políticas educativas que caracterizan el alto rendimiento o los países que muestran una gran mejora, entre otros. La propia OCDE también ofrece un análisis detallado de determinados países (en PISA 2012 estos fueron los Estados Unidos, Corea y Japón). Finalmente, cada país participante con su comunidad académica publica a menudo resultados más en profundidad comparativos y sobre el propio país en libros, artículos, informes y comunicados de prensa (véase, por

ejemplo, Instituto Nacional de Evaluación Educativa, 2014; Instituto Nacional para la Evaluación de la Educación, 2013; Ministerio de Educación de Chile, 2012, 2014; Prenzel, Sälzer, Klieme & Köller, 2013).

Interpretaciones intencionadas de los resultados de PISA, desde el punto de vista de los creadores de la prueba

La validez está estrechamente vinculada a los fines de una prueba, se establece para cumplir con el objetivo y las interpretaciones de las puntuaciones que un sistema de evaluación produce. De hecho, los Estándares plantean que "La validación lógicamente comienza con una declaración explícita del significado que se propone a las puntuaciones de la prueba, junto con una justificación de la pertinencia de la interpretación para el uso propuesto" (AERA, APA y NCME, 2014, p. 11). Por lo tanto, es necesario que prestemos atención a estas interpretaciones previstas en el marco del PISA. Como han demostrado algunos investigadores, los distintos agentes educativos interesados difieren en sus percepciones de los propósitos de una prueba y sus usos previstos, lo que hace difícil asumir un único conjunto de interpretaciones que puedan guiar el camino hacia usos apropiados (Linn, 1998; Taut et al., 2010). Sin embargo, de acuerdo a los Estándares (AERA, APA y NCME, 2014), el *desarrollador de la prueba* tiene la principal responsabilidad en la comunicación clara y pública de cuáles son las interpretaciones que se pretenden de la prueba diseñada (ver Estándares 1.1, 1.2, 1.5 y 1.6), y se hace responsable de presentar las evidencias de que estas interpretaciones y usos en realidad son válidas, y que los planteamientos subyacentes a ellos cuentan con respaldo empírico. En el caso de PISA, de la Organización para la Cooperación y Desarrollo Económico (OCDE) tiene esta función, y el resto de esta sección presenta qué documentos clave y páginas de internet, informan en este sentido.

Los materiales de comunicación por parte de la OCDE (por ejemplo, el folleto actual sobre PISA), subrayan las interpretaciones deseables y los usos para *toma de decisiones de política educativa* de los países participantes, en términos de (a) si el sistema educativo equipa a los jóvenes con importantes habilidades para la vida, también en comparación con otros países; (b) si el sistema educativo es justo; (c) si los jóvenes tienen potencial para el aprendizaje a lo largo de toda la vida (motivación, autoconfianza, etc.); y d) cómo evoluciona el rendimiento a lo largo del tiempo, se establecen metas de política y se evalúa el impacto de las decisiones sobre política educativa. Un video puesto de manera destacada en la página web de PISA explica que la finalidad de esa evaluación es mostrar a los países dónde están (también en relación con otros países) en la eficacia con la que educan a sus hijos, así como realizar un seguimiento de su progreso a lo largo del tiempo. Señala que el éxito en educación también incluye la manera en que se distribuye el rendimiento en educación de manera igualitaria entre los estudiantes en relación con los orígenes de los mismos. Por último, el vídeo habla de analizar las características de los sistemas educativos exitosos, mostrando lo que puede hacerse, así como las similitudes y diferencias entre los países, ayudando así a los países a revisar sus políticas en materia de educación y a diseñar otras nuevas y mejores. Por último, la sección de Preguntas Frecuentes de la web de PISA, afirma que PISA responde a los requerimientos por parte de los países miembros de "datos fiables y regulares sobre los conocimientos y las habilidades de sus estudiantes y el rendimiento de sus sistemas educativos", lo cual les permite " realizar un seguimiento de sus progresos en la consecución de los objetivos clave de aprendizaje".

El Informe Técnico PISA 2012 (OCDE, 2014, pág. 24) señala que cuando se vinculan los datos del rendimiento estudiantil con información contextual de los cuestionarios, PISA proporciona información para analizar las diferencias entre países en relación con los siguientes temas:

- Las relaciones entre factores a nivel de estudiante y rendimiento;
- Relación entre factores a nivel escolar y logro;
- Proporción de variación intra y entre las escuelas;
- La medida en que las escuelas moderan la relación entre factores a nivel individual y logro;
- Relación entre el contexto nacional y de los sistemas educativos y el logro;
- Los cambios en las relaciones antes mencionadas a lo largo del tiempo.

Además de puntuar en una escala con una media de 500 y una desviación estándar de 100, PISA también informa de los resultados en términos de seis niveles de conocimiento, en vistas a poder describir la alfabetización de estudiantes de manera más significativa. Descripciones resumidas de los seis niveles de pericia en la escala de alfabetización matemática, y cada subescala, figuran en el informe técnico (OCDE, 2014, pp. 297-301).

El Informe Técnico PISA 2012 (OCDE, 2014) no está estructurado de acuerdo a los Estándares de Pruebas Psicológicas y Educativas (AERA, APA y NCME, 2014) sino que, en cambio, se basa en estándares técnicos desarrollados específicamente para PISA 2012 (véase el anexo F). Según estos estándares, las "inferencias válidas en las comparaciones entre naciones" (p. 447) dependen de la consistencia, la precisión, la generalización y la prontitud con que se obtienen. Los estándares se dividen en estándares sobre datos, estándares sobre gestión y estándares sobre participación nacional.

Aunque sólo tangencialmente estén relacionados con el tema principal de este artículo, debemos mencionar que la validez transcultural es un tipo de evidencia de validez que se aborda explícitamente en el Informe Técnico PISA 2012. Los estándares sobre participación nacional se plantean para "asegurar que los instrumentos desarrollados a nivel internacional son ampliamente evaluados respecto a su validez trans-nacional, transcultural y trans-lingüística" (OCDE, 2014, p. 448). Además, la comparabilidad trans-cultural de las medidas en los cuestionarios de contexto de PISA son objeto de especial atención. El informe señala que "las diferencias transculturales en los estilos de respuesta se ha considerado que representan una grave fuente de sesgo en los estudios internacionales que utilizan escalas de tipo Likert" (OCDE, 2014, p. 53), y explica además que PISA 2012 se esforzó para hacer frente a esta amenaza con la validez, introduciendo nuevos formatos de items (viñetas de anclaje, detección de señales de corrección basadas en la sobrevaloración, la técnica de elección forzada de items y Tests de Juicio Situacional). Esta cuestión está más detallada en el capítulo sobre los procedimientos de escalado y construcción de la validación de los datos del cuestionario de contexto.

En resumen, Pisa pretende que sus resultados (tanto los basados en pruebas como los basados en cuestionarios) sean utilizados de al menos tres maneras distintas por sus principales destinatarios, a saber, *los elaboradores de políticas educativas* de los países participantes:

(1) Como *información de diagnóstico a nivel de país* (en términos de competencia en las áreas evaluadas, de equidad del sistema educativo, de otros factores a nivel individual, de la escuela y los niveles del sistema que están relacionados con los resultados de aprendizaje);

(2) Como *comparaciones a través del tiempo dentro de cada país* (lo que permite hacer un seguimiento de los progresos y evaluar el impacto de lasdecisiones políticas);

(3) Como *comparaciones con otros países* (para detectar las prácticas más exitosas y aprender de sus respectivos éxitos y fracasos).

Por lo tanto, es importante destacar que tanto las puntuaciones en los test como las mediciones *de constructos relevantes basadas en cuestionarios* deben considerarse para

validar interpretaciones y usos declarados de PISA como un programa de evaluación internacional.

Interpretaciones no deseadas de los resultados de PISA desde el punto de vista de los desarrollos de pruebas

Además de comunicar las interpretaciones planteadas de las puntuaciones de los test, los creadores de las pruebas tienen el deber de advertir explícitamente contra interpretaciones y usos no deseados, en los casos en los que ello sea posible debido a la experiencia previa en anticipar y abordar activamente este tema (AERA, APA y NCME, 2014, p. 19). De nuevo comprobamos lo que la OCDE, como desarrollador de las pruebas, informa en este sentido en los principales documentos y páginas de internet. El Informe Técnico de PISA 2012 no contiene ninguna mención de posibles interpretaciones no deseadas. Sin embargo, el vídeo explicativo de PISA mencionado anteriormente menciona en dos ocasiones: (a) que PISA no afirma que "Tal política produjo tal efecto"; (b) PISA no pretende crear una competición entre los sistemas educativos creando un ranking en términos de rendimiento en PISA. En el contexto de los rankings, la sección de Preguntas Frecuentes en el sitio web de PISA aclara que no es posible asignar un único puesto exacto en cada ámbito para cada país o jurisdicción participante. Asimismo, afirma que hay incertidumbre estadística en el muestreo de estudiantes involucrados y su extrapolación a la población, y que "por consiguiente, sólo es posible informar sobre un intervalo de posiciones (su rango superior y su rango inferior), dentro del cual un país puede ser colocado. Por ejemplo, en PISA 2003, Finlandia y Corea fueron presentados como ocupantes de los puestos 1º y 2º en PISA, cuando en realidad sólo podemos decir que, entre los países de la OCDE, el puesto de Finlandia estaba entre el 1º y el 3º y Corea estaba entre el 1º y 4º." Con respecto a la utilización de las clasificaciones y en respuesta a una carta abierta de Heinz-Dieter Meyer y Katie & Zahedi Zahedi (Meyer& Zahedi, 2014), la OCDE señala, "menos del 1% de los informes PISA está dedicado a las tablas sobre clasificaciones. El punto de vista de la OCDE es que debería corresponder a los países individuales hasta decidir en qué medida quieren ser comparados internacionalmente …" (véase el sitio web de Pisa, la sección de "Preguntas frecuentes").

A pesar de estos esfuerzos iniciales, aún parece justo decir que en los textos de comunicación de la OCDE sobre PISA se puede encontrar sobre interpretaciones infundadas y usos potencialmente dañinos de los resultados de PISA en los países participantes y a nivel internacional, con la única excepción de los rankings.

Evidencia empírica existente sobre la validez de las consecuencias de PISA

La literatura[i] sobre los usos y consecuencias de PISA se centra principalmente en las tres primeras oleadas de PISA (2000, 2003 y 2006). Este cuerpo de literatura incluye algunos estudios que han revisado los efectos sobre las políticas de PISA en diferentes países, especialmente en los países europeos. Por ejemplo, Baird et al. (2011) examinaron la respuesta política a PISA en seis países/regiones, sobre cómo participantes de alto rendimiento (Canadá y Shanghai-China) se compararon con los países europeos que puntúan generalmente alrededor de la media (Inglaterra, Francia, Noruega y Suiza), pero en los cuales había habido un interesante político de PISA. Otro ejemplo fue el de una evaluación externa del impacto político de PISA, encargado por el Consejo de Administración (OCDE, 2008)., el cual utilizaba tanto métodos cuantitativos como cualitativos para evaluar la relevancia, la eficacia y la sostenibilidad, añadiendo también los impactos inesperados de PISA. En la vertiente cuantitativa, un conjunto de agentes sociales (legisladores, funcionarios del gobierno local, directores de escuela, padres de familia, académicos y representantes de los medios de comunicación) de 43 países y economías fueron encuestados mediante correo electrónico. En la vertiente cualitativa,

diferentes grupos de agentes sociales implicados participaron en entrevistas y en grupos focales. Este estudio también incluyó estudios de casos en cinco países y economías (Canadá, Hong Kong-China, Noruega, Polonia y España), considerando sus diferencias en términos de rendimiento en PISA, el impacto sobre su política, la equidad y la estructura de gobierno. Por último, unos pocos estudios de (Breakspear, 2012; Martens, Nagel, Windzio, & Weymann, 2010), entrevistaron a representantes y expertos de los países de la OCDE, y también analizaron los documentos de política, a fin de poner de relieve la diversidad de las respuestas nacionales a la publicación de los resultados de PISA.

Estos estudios pusieron de relieve una serie de resultados: (a) el impacto de PISA fue mayor a nivel nacional que a nivel regional o escolar; b) los que toman las decisiones políticas fueron identificados como el grupo de interesados más relevante; (c) los países valoraban cada vez más los conocimientos evaluados en PISA; (d) PISA se utilizaba habitualmente para monitorear tanto el desempeño de un país como la equidad; (e) la influencia de PISA sobre la política parecía ir en aumento a lo largo del tiempo; (f) PISA tiene el potencial para 'definir' retos de la política educativa y establecer la agenda para el debate político en los niveles nacional y estatal; (g) la mayoría de los países han adoptado algún tipo de reforma política o iniciativa, en diversa medida, principalmente en función de su nivel de rendimiento - como respuesta directa a PISA, en algún momento entre las diversas oleadas de la encuesta (Baird et al., 2011; Breakspear, 2012; OCDE, 2008)..

Basándonos en la literatura mencionada anteriormente, así como de otras fuentes, la evidencia sobre la validez de las consecuencias de PISA se ha clasificado respecto a las interpretaciones intencionales y no intencionales y los usos enumerados en las secciones anteriores.

(1a) información de diagnóstico a nivel de país

En este grupo, nos centraremos en los ejemplos de Francia y Alemania. En el caso francés, su rendimiento en lectura ha estado consistentemente alrededor de la media de la OCDE (Urteaga, 2010).. Sin embargo, la amplia difusión de la distribución de la puntuación de los estudiantes en lectura ha suscitado la preocupación de responsables políticos. En otras palabras, los resultados muestran que hay una gran proporción de estudiantes tanto en la banda superior de rendimiento como en la banda inferior y, además, esta tendencia fue más pronunciada en 2009 que en 2000. En respuesta, el gobierno francés ha anunciado una serie de reformas en el curriculum de la escuela primaria, tales como la introducción de una estrategia de lucha contra el analfabetismo, pero también el apoyo al aprendizaje personalizado a lo largo de todo el sistema para ayudar a los alumnos de bajo rendimiento, complementado con la mayor autonomía de las escuelas para poder administrar sus propios presupuestos. (Baird et al., 2011).

Los resultados alemanes en PISA 2000 mostraron que el rendimiento en ciencias de los estudiantes alemanes ciencia logro fue significativamente inferior a la media de la OCDE, y que su rendimiento en lectura y matemáticas fue similar. Estos resultados fueron mucho menores de lo esperado ('PISA-Schock") y, por consiguiente, significó una noticia devastadora para el sistema educativo alemán, anteriormente considerado uno de los mejores del mundo. Como resultado, se introdujeron importantes cambios en el sistema educativo alemán. En primer lugar, los cambios en el discurso político fueron acompañados con un amplio programa de reformas que incluía una serie de iniciativas (por ejemplo, programas para mejorar la calidad de la enseñanza y el aumento de la financiación de las escuelas), pero lo que es más importante, la introducción de los Estándares Nacionales de Educación (NES). El núcleo de la reforma fue el concepto de habilidades y competencias dominando la noción tradicional alemana de educación. Por

ejemplo, la NES describe las competencias científicas que se espera que los estudiantes hayan adquirido al final de su educación secundaria inferior (Neumann, Fischer, & Kauertz, 2010).. Además, en términos de los procesos de desarrollo curricular, el control de resultados y la evaluación externa han adquirido una importancia fundamental. Finalmente, el discurso académico, cambió su dirección hacia un mayor énfasis en la investigación empírica de prácticas pedagógicas (Ertl, 2006).

(1b) comparaciones a lo largo del tiempo dentro de cada país

En esta categoría encontramos los casos de Polonia y Hong Kong (China). En el primer caso, las importantes mejoras de Polonia en los resultados de PISA se han relacionado con una mejora significativa de dos períodos (2000-2003 y 2009-2012) (Amoroso et al., 2015). Este escenario positivo, especialmente la mejora en el periodo 2000-2003, ha sido atribuido a las reformas educativas en Polonia durante los años noventa, aunque es difícil separar los efectos concretos de cada elemento de la reforma. Esta reforma fue diseñada no sólo para aumentar las oportunidades educativas para todos los estudiantes, sino también para superar el sistema educativo heredado de la época comunista. La mencionada reforma incluía cambios estructurales como demorar la selección entre itinerarios de enseñanza general y profesional por un año, y la introducción de tres años de educación general secundaria obligatoria (*gimnazjum*), con efectos positivos sobre los resultados de rendimiento (Jakubowski, 2015). Se sugirió una hipótesis acerca de la relación entre la reforma educativa polaca y su mejora en PISA, ya que los cambios en el currículum y la evaluación de los estudiantes llevaron a la mejora de las destrezas cognitivas, de tal modo que los alumnos que fueron evaluados en 2012 ya habían completado tres años de secundaria inferior bajo el nuevo currículum (Amoroso et al., 2015; Jakubowski, 2015). En resumen, a pesar de que PISA no fue el impulsor del cambio en las políticas de educación de Polonia, fue utilizado como una herramienta de seguimiento a fin de controlar la evolución de las calificaciones de los estudiantes en términos de las repercusiones en la política.

Como en el caso de Polonia, en Hong Kong (China) PISA 2003 y 2006 resultados no fueron identificados como el impulsor clave para las reformas, sino que sirvió como un importante instrumento de vigilancia, aunque la mejora en el rendimiento se atribuye al programa de reforma educativa (OCDE, 2008). Esta reforma incluyó un nuevo curriculum en 2002 y, posteriormente, un nuevo sistema de seguimiento del estudiante. Además, el Estudio de la OCDE (2008) sugiere que Hong Kong utilizó PISA como guía para la construcción de sus nuevos objetivos educativos, sustituyendo el énfasis a los estudiantes sobre la adquisición de conocimientos por el desarrollo de la comprensión, la resolución de problemas, el razonamiento y el pensamiento estratégico. Finalmente, Hong Kong introdujo una amplia gama de reválidas o exámenes de rendimiento (a los 6, 9 y 15 años de edad) y la adopción de buenas prácticas internacionales para intervenir sobre las tasas de deserción escolar, las tasas de finalización de la secundaria superior y la participación en el aprendizaje a lo largo de la vida.

(1c) Las comparaciones con otros países

Una consecuencia importante de los estudios internacionales como PISA es que los políticos y los encargados de formular políticas tienen que responder de la posición de sus países en los rankings. Esto es particularmente relevante cuando el resultado de un país es peor de lo esperado, ya sea bajando en el ranking o obteniendo peores resultados que los países vecinos (Stobart & Eggen, 2012). Un ejemplo interesante de esto último es el caso de Noruega. En 2000 y 2003, los resultados de PISA noruegos estaban por debajo de la media de la OCDE y, lo que es más importante, también por debajo de sus vecinos escandinavos (Suecia, Dinamarca y Finlandia), a pesar de una buena financiación y de la buena auto-confianza en su sistema

educativo (OCDE, 2008). Por consiguiente, la desfavorable comparación con sus vecinos en PISA ha tenido un impacto significativo sobre la política educativa en Noruega, impulsando a una serie de reformas tanto en términos de evaluación como de políticas sobre el currículum (Baird et al., 2011; Chung, 2016; Elstad, 2010).. El 'Norwegian PISA shock" se convirtió en una fuerza impulsora para la reforma del sistema educativo, que incluía cambios en los niveles primario y secundario: un nuevo currículum, con mayor énfasis en resultados mensurables; proyectos gubernamentales para promover de manera amplia la evaluación formativa; un nuevo sistema de evaluación de calidad con pruebas nacionales; nuevas regulaciones para los exámenes y para que los profesores den información de las calificaciones finales (Tveit,, 2013).

(2) consecuencias no intencionadas

Basándose en las respuestas a las encuestas de diferentes grupos de agentes educativos implicados, procedentes de 43 países y economías, la evaluación externa de la OCDE (2008) describió efectos no intencionados de PISA, tanto positivos como negativos. Respecto a los impactos inesperados positivos, los hallazgos incluyen niveles particularmente altos de interés del público y los debates a partir de los resultados de PISA; la asignación de un mayor valor a las habilidades evaluadas por PISA y el ajuste de las evaluaciones nacionales a este objetivo; el aumento de la colaboración entre los distinto grupos de agentes implicados para mejorar los resultados de su país y el sistema educativo y el creciente interés en la investigación educativa empírica. En el lado negativo, el informe menciona las discusiones nacionales que tratan de la responsabilidad por el bajo rendimiento en grupos particulares (por ejemplo, maestros), resultando en una "cultura de la culpa" y la utilización de PISA para la legitimación de las reformas educativas que de otro modo serían más abiertamente discutidas y protestadas.

En términos de pruebas específicas de cada país encontramos referencias sobre Turquía, Japón, España y Chile. En el caso de Turquía, sus resultados en PISA 2003 y 2006 eran inferiores a la media de la OCDE. La reacción de los educadores, legisladores y periodistas se concentró en los pobres resultados en comparación con otros países (los puestos obtenidos en la tabla de ordenación de países). Por ejemplo, los periódicos se interesaron principalmente en los rankings, ignorando así otra información relevante que revelaron los resultados de PISA y lo que implicaban para la mejorar del sistema educativo (Gür, Çelik, & Özoğlu, 2012). Gür y sus colegas (2012) examinaron los documentos públicos (por ejemplo, informes oficiales y boletines informativos publicados por el Ministerio de Educación) y llegaron a la conclusión de que las autoridades ya habían decidido introducir una nueva reforma educativa mucho antes de que se publicaran los resultados de PISA 2003. Sin embargo, los funcionarios del gobierno utilizaronn los resultados de PISA para justificar la necesidad de una reforma del sistema educativo sin un cuidadoso examen de los resultados y lo que significaban para el sistema en su conjunto.

Asimismo, a pesar de que Japón mostraba unos buenos resultados en PISA 2000, sus resultados en PISA 2003 fueron interpretadas por la prensa como una tendencia a la baja, lo que se tradujo en una percepción de 'crisis' que alentaron un considerable debate político y público sobre la reforma de la educación. En respuesta a la disminución en las puntuaciones, el gobierno japonés cambió una polémica política de un curriculum de baja presión en favor de las prácticas de evaluación nacional (Takayama, 2008).. Sin embargo, desde un punto de vista objetivo los resultados en PISA 2003 no fueron estadísticamente diferentes de PISA 2000 en las matemáticas; sólo hubo una disminución estadísticamente significativa en lectura, que en verdad representa un punto débil de los estudiantes japoneses conocido desde hace tiempo. Además, interpretación de la prensa japonesa

sobre clasificaciones no mencionaban que los que mejores puntuaciones obtuvieron en 2003 (los Países Bajos y Hong Kong) no habían sido incluidos en PISA 2000 (Takayama, 2008)..

Asimismo, en el año 2013 España presentó una serie de reformas educativas explícitamente inspirada y justificada por los pobres resultados de PISA en 2012 (Choi & Jerrim, 2015; OCDE, 2014). En particular, la última y más importante iniciativa fue la Ley Orgánica para la Mejora de la Calidad Educativa (LOMCE, en español), que incluye iniciativas tales como una mayor autonomía para las escuelas, nuevas pruebas de diagnóstico preventivo en educación primaria (en sexto curso), más itinerarios de formación profesional a partir de los últimos años de la enseñanza secundaria inferior y exámenes al final de los estudios de educación secundaria inferior y superior (OCDE, 2014). Sin embargo, algunos autores han argumentado que las interpretaciones de los resultados de PISA utilizados para legitimar esta reforma han sido incorrectas, en particular los relacionados con los puestos ocupados por España en el ranking y las comparaciones con vecinos europeos (Choi & Jerrim, 2015; Bonal & Tarabini, 2013). Esto ha sido cierto para los políticos (Jornet, 2013, 2016c) y la prensa (Carabaña, 2008).. Según estos autores, esto pone de manifiesto las consecuencias negativas que vienen desde el uso exclusivo e inexacto de las clasificaciones para la formulación de políticas educativas (Jornet, 2016a, 2016b).

Por último, Chile experimentó una mejoría significativa en la lectura durante el periodo 2000-2009, especialmente entre los estudiantes de bajos ingresos, reduciendo la brecha de logros por cuestiones socioeconómicas (OCDE, 2010). No obstante, los encargados de la formulación de políticas educativas y la prensa optó por ignorar esta mejora y exclusivamente se fijó en el *vaso medio vacío*, destacando la posición de Chile por debajo de la media, en comparación con los países de la OCDE y principalmente, subrayaron la necesidad de continuar la reforma educativa (Ravela, 2011). Aunque en las comparaciones internacionales a lo largo del tiempo Chile es uno de los países que más han progresado durante la última década, lo que debería haber sido una buena noticia a nivel nacional fue malinterpretado con fines políticos.

En resumen, según la literatura sobre diferentes ejemplos de casos, los resultados de PISA desencadenaron importantes reacciones en la dirección pretendida en algunos países. Éstos generalmente parecen ser países en los que PISA diagnosticó un rendimiento que quedaba fuertemente por debajo de las expectativas nacionales -que podría estar basados en un alto rendimiento anterior o en expectativas demasiado ambiciosas no cumplidas, o en una comparación desfavorable con otros “países semejantes". A veces, los resultados de PISA sobre diferentes temas se complementaron con información sobre la distribución por categorías de desempeño e indicadores de equidad educativa en estos países. Asimismo, las tendencias reveladas a lo largo del tiempo sirvieron como información de seguimiento para evaluar los progresos realizados en relación con determinados tipos de habilidades y ámbitos. En términos de usos no pretendidos, estos parecen haber sido emprendidos por actores políticos nacionales, por razones políticas, a veces apoyados por los medios de comunicación, y en su mayoría relacionados con el uso de clasificaciones de PISA para generar un sentido de urgencia y para legitimar las reformas educativas, estableciendo vínculos de causalidad directa de ciertas políticas nacionales con los resultados de PISA, además comparaciones injustas. Esto ha tenido consecuencias negativas en cuanto al diagnóstico de los fallos de las políticas educativas por un lado, o empujando a realizar reformas injustificadas, por el otro.

Investigación pendiente sobre la validez de las consecuencias de PISA

De acuerdo a los Estándares de Pruebas Psicológicas y Educativas (AERA, APA y NCME, 2014), el desarrollador de la prueba -

en este caso, la OCDE - es responsable de presentar "argumentos lógicos o teóricos y evidencia empírica" (p. 24) que apoyen las interpretaciones y usos de los resultados de PISA que sugiere, explícita o implícitamente, el desarrollador de la prueba. Los Estándares van tan lejos como para incluir cualquier beneficio indirecto que se anticipe en base al programa de evaluación (véase Estándar 1.6). Esto interpretaciones y usos intencionados corresponden a las presentadas en la secciones anteriores. Aunque el Informe Técnico contiene una riqueza de información que es relevante para aportar un argumento de validez para PISA, el informe no presenta esta información organizada de acuerdo a las exigencias de la validez, o respecto a las interpretaciones y usos pretendidos, ni su web contiene otros documentos que traten específicamente de la validación de PISA, ni presenta criterios de validez para cada una de sus interpretaciones y usos. Como se mencionó anteriormente, hubo una evaluación externa sobre el impacto político de PISA (OCDE, 2008), pero este estudio no estaba técnicamente orientado (es decir, con mediciones) ni presentaba argumentos y pruebas integrales para cada interpretación y uso previsto. Proporcionaba evidencias de los usos intencionales de PISA a nivel de país, y se estudiaron también explícitamente los efectos no intencionales de PISA; cuyos resultados han sido presentados más arriba. Sin embargo, la investigación global de la validez de PISA que esté impulsada por sus interpretaciones y usos intencionales, tal y como es exigido por los Estándares de medición ampliamente aceptados (AERA, APA y NCME, 2014) parece seguir siendo una tarea pendiente, al menos según la documentación disponible públicamente en el sitio web de PISA. Sin embargo, PISA presta considerable atención a la validez transcultural, un aspecto no incluido en otros marcos de estándares citados anteriormente.

Del mismo modo, aunque se aconseja cierta cautela en el uso de los rankings, la información sobre interpretaciones y usos no intencionales o no fundamentadas está prácticamente ausente en la web de PISA y en la documentación contenida en la misma. Como los Estándares (AERA, APA y NCME, 2014) señalan, el desarrollador de la prueba no se hace responsable de los usos y consecuencias no pretendidas, a menos que se deba a defectos en el propio test (una representación insuficiente del constructo o una varianza irrelevante para el constructo). Además, las instituciones locales coordinadoras de PISA asumen una mayor responsabilidad en este sentido, pero ni siquiera ellas pueden evitar que los responsables políticos y los medios lleguen a conclusiones erróneas sobre la base de los resultados de PISA. Sin embargo, como se señala también en las recomendaciones de la evaluación externa de PISA (OCDE, 2008), la OCDE y sus colaboradores podrían hacer más para llamar la atención sobre las prácticas disparatadas, para propiciar el desarrollo de una alfabetización sobre la evaluación y para ser activos en la promoción de usos previstos a la vez que identifican, y advierten, sobre usos no pretendidos.

Este escenario sobre validación pendiente es sorprendente, habida cuenta de los considerables recursos financieros que requiere PISA y el alto nivel de conocimientos técnicos de los que participan habitualmente en su desarrollo y análisis. Los países participantes pueden exigir pruebas de esa índole en el futuro (por ejemplo, véase Martínez Rizo et al., 2015; Schafer, Wang y Wang, 2009; Taut, Santelices & Stecher, 2012) y la OCDE podría dedicar un capítulo específico en el Informe Técnico (y en la sección correspondiente de la web) para presentar evidencia que apoye las interpretaciones y usos propuestos, así como indicar en otros lugares, igualmente visibles, advirtiendo contra los no admitidos o no intencionados.

Conclusiones

"La teoría sobre validez es rica, pero la práctica de la validación a menudo está empobrecida" (Brennan, 2006, p. 8). Esta conclusión, que a menudo se expresa en

grupos especializados en medición educativa, parece aplicarse también al sistema de pruebas de PISA, y particularmente al aspecto consecuencial de la validez. Por encima de todo, los desarrolladores de pruebas deben ser responsables de validar interpretaciones, usos y consecuencias, y dicha documentación deberá estar a disposición del público de manera oportuna. Tales evidencias difícilmente serán completas y definitivas, sino que son más bien un esfuerzo explícito que involucra recursos sustanciales y que deben estar visibles.

Los creadores de pruebas tienen una responsabilidad decreciente sobre los usos de las pruebas, más allá de los resultados de las pruebas (y los correspondientes datos del cuestionario) que producen. Aunque pueden ser responsabilizados por el tipo de interpretaciones que ellos plantean, así como de comunicar claramente cómo deben y no deben usarse las puntuaciones de sus pruebas, en realidad impedir los usos inapropiados y consecuencias negativas está claramente fuera de su esfera de influencia; sin embargo, los creadores de pruebas pueden desempeñar una función activa en la educación de los usuarios de evaluaciones en el uso apropiado de las pruebas y en llamar la atención del público respecto a malas interpretaciones previsibles y a casos reales de mal uso de los datos. Este artículo no pretende juzgar cuánto ha hecho PISA y cuánta responsabilidad tiene en impedir cualquier abuso que se haya producido a nivel nacional en el pasado. De hecho, la evaluación externa sobre los impactos políticos, intencionales y no intencionales, de PISA (OCDE, 2008) incluyen dos recomendaciones definitivas en este sentido: a) "como mínimo, PISA debería elaborar directrices de difusión para quienes participan en el programa"; y b) "PISA debería considerar, como mínimo, la creación de un grupo sobre políticas para los países que soliciten su asesoramiento en la creación de actuaciones para una mejor utilización de los resultados de PISA" (p. 9). Un buen ejemplo son las directrices para la utilización de las pruebas basadas en PISA para escuelas (OCDE, 2013). En cualquier caso, agentes nacionales, tales como los organismos evaluadores, ministerios y académicos también desempeñan un papel clave para influir en que los resultados de PISA sea adecuadamente interpretados y utilizados en sus respectivos países.

Referencias

American Educational Research Association, American Psychological Association & National Council for Measurement in Education [AERA, APA & NCME] (2014). *The Standards for Educational and Psychological Testing. Washington*, D.C.: AERA.

Amoroso, J. M., Moreno, J. M., Gortazar, L., Herrera Sosa, K. M., Kutner, D., & Bodewig, C. (2015). *Poland - Skilling up the next generation: an analysis of Poland's performance in the program for international student assessment*, 1–21. Disponible en http://documents.worldbank.org/curated/en/2015/12/25518729/poland-skilling-up-next-generation-analysis-poland%E2%80%99s-performance-program-international-student-assessment

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Disponible en http://research-information.bristol.ac.uk/en/publications/policy-effects-of-pisa(833739c4-7e0a-4c18-b249-a3f12120065f).html

Bonal, X., & Tarabini, A. (2013). The role of PISA in shaping hegemonic educational discourses, policies and practices: The case of Spain. *Research in Comparative and International Education*, *8*(3), 335–341. http://dx.doi.org/10.2304/rcie.2013.8.3.335

Breakspear, S. (2012). The policy impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. *OECD Journals*, (71), 1–32. DOI: http://dx.doi.org/10.1787/19939019

Brennan, R. (2006). Perspectives on the Evolution and Future of Educational

Measurement. En R. Brennan (ed.), *Educational Measurement*, 4th ed., pp. 1-16. Westport, CT: Praeger.

Carabaña, J. (2008). *Las diferencias entre países y regiones en las pruebas PISA*. Madrid: Colegio Libre de Eméritos

Choi, A., & Jerrim, J. (2015). The Use (and Misuse) of PISA in Guiding Policy Reform: The Case of Spain. *SSRN Electronic Journal*, 1–16. DOI: http://dx.doi.org/10.2139/ssrn.2580141

Chung, J. (2016). The (mis)use of the Finnish teacher education model: "policy-based evidence-making"? *Educational Research*, *58*(2). DOI: http://dx.doi.org/10.1080/00131881.2016.1167485

Cronbach, L. (1988). Five perspectives on the validity argument. In H. Wainer & H. Braun (eds.), *Test validity*, pp. 3-17. Hillsdale, NJ: Lawrence Erlbaum.

Elstad, E. (2010). P*isa Debates and Blame Management Among the Norwegian Educational Authorities*: Press Coverage and, *48*, 10–22.

Ertl, H. (2006). Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*(5), 619–634. DOI: http://dx.doi.org/10.1080/03054980600976320

Gür, B. S., Çelik, Z., & Özoğlu, 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

Instituto Nacional de Evaluación Educativa. (2014). *PISA 2012 Resolución de problemas de la vida real. Resultados de matemáticas y lectura por ordenador. Informe Español. Versión preliminar*. Instituto Nacional de Evaluación Educativa. Disponible en http://www.mecd.gob.es/dctm/inee/internacional/pisa2012-resolucionproblemas/pisaresoluciondeproblemas.pdf?documentId=0901e72b8198bee8

Instituto Nacional para la Evaluación de la Educación. (2013). *México en 2012*. México: Instituto Nacional para la Evaluación de la Educación. Disponible en http://www.inee.edu.mx/images/stories/2013/principal/PISA2013/PISA_2012041213web1.pdf

Jakubowski, M. (2015). Opening up opportunities: education reforms in Poland, (January).

Jornet, J. (2013, Enero 30). Cuestionados los supuestos malos datos españoles del informe Pisa. *Comunidad Valenciana*. Valencia. Disponible en http://ccaa.elpais.com/ccaa/2013/01/30/valencia/1359572336_318312.html

Jornet, J. (2016a). *España en PISA*. Valencia: Ateneo Mercantil de Valencia.

Jornet, J. (2016b, Enero 26). La educación no está tan mal; el informe PISA. *Levante. El Mercantil Valenciano*, p. 10. Valencia.

Jornet, J. (2016c, Enero 26). Cómo desmontar el informe PISA. *Las Provincias*. Valencia. Disponible en http://www.lasprovincias.es/comunitat/20160126/como-desmontar-informe-pisa-20160126001834-v.html

Kane, M. (2006). Validity. In Brennan, R. (ed.), *Educational Measurement*, 4th ed., pp. 17-64. Westport, CT: Praeger.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448-457.

Linn, R. (1998). Partitioning responsibility for the evaluation of the consequences of use. *Educational Measurement: Issues and Practice*, 17(2), 28-30.

Martens, D. K., Nagel, A.-K., Windzio, M., & Weymann, A. (2010). *Transformation of Education Policy*. Basingstoke: Palgrave.

Martinez Rizo, F. (2015). *Las pruebas ENLACE y Excale. Un estudio de validación. Cuaderno de Investigación No. 40*. México. DF: Instituto Nacional para la Evaluación de la Educación. Disponible en http://publicaciones.inee.edu.mx/buscadorPub/P1/C/148/P1C148.pdf

Mehrens, W. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16-18.

Messick, S. (1989). Validity. In R. Linn (ed.), *Educational measurement* (3rd ed.), pp. 13-103. Washington, D.C.: American Council on Education.

Meyer, H.-D. & Zahedi, K. (2014, May 4). *Open Letter to Andreas Schleicher*, OECD, Paris. Disponible en http://www.globalpolicyjournal.com/blog/05/05/2014/open-letter-andreas-schleicher-oecd-paris

Ministerio de Educación de Chile. (2012). *Evidencias para Políticas Públicas en Educación: Selección de Investigaciones Concurso Extraordinario FONIDE-PISA*. Santiago de Chile: Ministerio de Educación de Chile. Disponible en https://s3.amazonaws.com/archivos.agenciaeducacion.cl/documentos-web/Estudios+Internacionales/PISA/Evidencias_para_Politicas_Publicas_en_Educacion_FONIDE_PISA.pdf

Ministerio de Educación de Chile. (2014). *Informe Nacional Resultados Chile Pisa 2012*. Santiago de Chile: MINEDUC. Disponible en https://s3.amazonaws.com/archivos.agenciaeducacion.cl/documentos-web/Estudios+Internacionales/PISA/Informe_Nacional_Resultados_Chile_PISA_2012.pdf

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: the impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, *8*(3), 545–563. http://doi.org/10.1007/s10763-010-9206-7

Organisation for Economic Co-Operation and Development (2008). *External evaluation of the policy impact of PISA*, (November), 3–5. Disponible en http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/PISA/GB(2008)35/REV1&docLanguage=En

Organisation for Economic Co-Operation and Development (2010). PISA 2009 Results: Executive Summary. *Executive Summary*, 1–21. Disponible en http://www.oecd.org/pisa/pisaproducts/46619703.pdf

Organisation for Economic Co-Operation and Development (2013). *General guidelines for the availability and uses of the PISA-based test for schools*. Disponible en https://www.oecd.org/pisa/aboutpisa/PISA-based-test-for-schools-guidelines.pdf

Organisation for Economic Co-Operation and Development (2014, April). *Education Policy Outlook Spain*. Disponible en http://www.oecd.org/edu/EDUCATION%20POLICY%20OUTLOOK%20SPAIN_EN.pdf

Organisation for Economic Co-Operation and Development (2014). *PISA 2012 Technical Repor*t. Disponible en https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Popham, W. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9-13.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (eds.) (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland (PISA 2012. Improvements and challenges in Germany)*. Münster: Waxmann.

Ravela, P. (2011). ¿Qué hacer con los resultados de PISA en América Latina? *PREAL. Programa de Promoción de La Reforma Educativa En América Latina y El Caribe*, *58*.

Schafer, W., Wang; J. & Wang, V. (2009). Validity in action: State assessment validity evidence for compliance with NCLB. En R. Lissitz (ed.), *The concept of validity*, pp. 173-193. Charlotte, NC: Information Age Publishing.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5-8.

Stobart, G., & Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1–6. DOI: http://dx.doi.org/10.1080/0969594X.2012.639191

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

Taut, S., Santelices, V. & Stecher, B. (2012). Validation of a national teacher assessment and improvement system. *Educational Assessment Journal*, *17*(4), 163-199. DOI: http://dx.doi.org/10.1080/10627197.2012.735913

Taut, S., Santelices, V., Araya, C. & Manzi, J. (2010). Theory unterlying a national teacher evaluation program. *Evaluation and Program Planning*, *33*, 477-489. DOI: http://dx.doi.org/10.1016/j.evalprogplan.2010.01.002

Tveit, S. (2013). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 221–237. DOI: http://dx.doi.org/10.1080/0969594X.2013.830079

Urteaga, E. (2010). Los resultados del estudio PISA en Francia. *Revista Complutense de Educación*, *21*, 231–244.

Notas

[i] En el presente trabajo se utilizaron diversas herramientas de búsqueda y bases de datos, incluyendo Web of Science, ScienceDirect, Scopus y Google Scholar para buscar los estudios potencialmente relevantes. Las palabras clave para esta búsqueda fue PISA y sus combinaciones con los siguientes términos*: interpretaciones, efectos, usos, consecuencias, toma de decisiones, la validez, la validación, la consecuente validez.* A continuación, incluimos la combinación de *los resultados de PISA* con el mismo conjunto de términos. Además, repetimos el mismo proceso en español.

Apéndice A

(Tomado de Martínez Rizo, 2015)

USOS Y CONSECUENCIAS

1. Se presentan argumentos lógicos o teóricos y evidencia empírica que respalde los usos y consecuencias previstas, y se evita sugerir otros si no tengan suficiente apoyo teórico o empírico.
2. Se documenta y evalúa el grado en que se producen las consecuencias previstas y/o deseables de la prueba.
3. Los resultados de las pruebas se reportan en plazos razonables y se proveen mecanismos de difusión y acceso para distintos usuarios, sin discriminación.
4. Se apoya a instituciones y usuarios para desarrollar la capacidad necesaria para la adecuada interpretación y utilización de los resultados.
5. Se informa a los usuarios sobre los propósitos y características de la prueba, lo que puede o no medir y los usos y consecuencias previstas. Se ofrecen ejemplos e información suficiente sobre la adecuada interpretación de los resultados.
6. Se utiliza lenguaje claro y preciso sin jerga técnica innecesaria; se explican términos técnicos en lenguaje claro y comprensible.

7. Se ofrece marco normativo para evaluar el desempeño de los examinados. Se describen el perfil y características de la población de referencia.
8. Se da información para minimizar posibilidad de interpretaciones incorrectas. Se notan limitaciones y errores comunes al comparar años, dominios, grupos o niveles de agregación. Se usan categorías precisas que no estigmaticen.
9. Se advierte sobre usos para los que no existe suficiente evidencia de validez. Si bien no pueden preverse todos los usos o interpretaciones inapropiadas, se busca identificar y acotar los más comunes.
10. Se documenta la existencia de usos o consecuencias imprevistas, ya sean adecuadas y positivas, negativas o/inadecuadas.
11. Cuando existe evidencia confiable de estos usos inapropiados se investigan en grado y detalle adecuado. Si persisten estos usos se informa a los usuarios y se intenta tomar acciones correctivas.

Autores / Authors	To know more / Saber más
Taut, Sandy (staut@uc.cl). Licenciada en Psicología por la Universidad de Colonia (Alemania), obtuvo su doctorado en educación por la universidad de California Los Angeles (UCLA, USA). Es profesora asistente en la Facultad de Psicología de la Pontificia Universidad Católica de Chile e investigadora asociada en el Centro de Medición MIDE UC. Es la autora de contacto para este. Su dirección es: Pontificia Universidad Católica de Chile, Escuela de Psicología, Centro de Medición MIDE UC, Avda Vicuña Mackenna 4860, Macul, Santiago (Chile).	Google académico ResearchGate academia.edu
Palacios, Diego (dfpalaci@uc.cl). Investigador asociado en el Centro de Medición MIDE UC. En la Pontificia Universidad Católica de Chile. Su dirección postal es: Pontificia Universidad Católica de Chile, Escuela de Psicología, Centro de Medición MIDE UC, Avda Vicuña Mackenna 4860, Macul, Santiago (Chile).	Google académico ResearchGate academia.edu

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]