Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Capacidad evaluativa, validez cultural y validez consecuencial en PISA

Assessment Capacity, Cultural Validity and Consequential Validity in PISA

Solano-Flores, Guillermo [(1)] **& Milbourn, Tamara** [(2)]

(1) Stanford University. (2) University of Colorado Boulder.

Resumen

Las evaluaciones internacionales de estudiantes han desempeñado un papel cada vez más importante en la política educativa. Estas comparaciones internacionales basadas en pruebas generan información valiosa sobre el rendimiento del estudiantado de cada país participante y los factores sociales y contextuales asociados. Una imagen compleja de los factores culturales, económicos y sociales que dan forma a la participación de PISA empieza a emerger. Nuestro objetivo es entender la relación entre la capacidad evaluativa nacional y la forma en que los países participan en estas comparaciones internacionales. Proponemos un marco conceptual para examinar la capacidad evaluativa como clave para abordar dos aspectos de la validez: cultural y consecuencial. Asimismo, se discuten las múltiples facetas de la capacidad evaluativa como condiciones para abordar la validez cultural y validez consecuencial en comparaciones internacionales.

Fecha de recepción
04 Abril 2016

Fecha de aprobación
20 Junio 2016

Fecha de publicación
21 Junio 2016

Palabras clave:

PISA; capacidad evaluativa; validez cultural; validez consecuencial

Abstract

International student assessments have played an increasing important role in educational policy. These international test comparisons generate valuable information about each participating country's student performance and the social and contextual factors. A complex picture of the cultural, economic, and social factors that shape PISA participation is emerging. We aim to understand the relationship between national assessment capacity and how countries participate in international test comparisons. We propose a framework for examining assessment capacity as key to addressing two aspects of validity -cultural and consequential. Also, we discuss the multiple facets of assessment capacity as conditions for addressing cultural validity and consequential validity in international test comparisons

Reception Date
2016 April 04

Approval Date
2016 June 20

Publication Date:
2016 June 21

Keywords:

PISA; assessment capacity; cultural validity; consequential validity

Las pruebas internacionales, tales como TIMSS (Estudio de Tendencias en Matemáticas y Ciencias), PISA (Programa para la Evaluación Internacional de Alumnos) y PIRLS (Estudio Internacional del Progreso en Comprensión Lectora) son cada vez más importantes en la política educativa. Estas pruebas comparativas internacionales generan información valiosa sobre el desempeño de estudiantes en cada país participante y los factores sociales y contextuales que pueden explicar las diferencias en ese desempeño

Autor de contacto / Corresponding author

Solano-Flores, Guillermo. Stanford University Graduate School of Education. 485 Lasuen Mall. Stanford, CA 94305-3096. United States. gsolanof@stanford.edu

(p.ej. infraestructura, características del profesorado y del currículum). Estos datos pueden potencialmente ayudar a las jurisdicciones participantes a tomar decisiones informadas en sus políticas educativas. En efecto, la OECD (Organización para la Cooperación y el Desarrollo Económico, por sus siglas en inglés) alienta a los países participantes y sus economías asociadas a desarrollar nuevas políticas educativas basándose en las fortalezas y debilidades identificadas:

> PISA ofrece una visión de la política y la práctica educativa y ayuda a monitorear las tendencias en la adquisición de conocimientos y habilidades de los estudiantes en todos los países y en diferentes subgrupos demográficos de cada país. Los hallazgos permiten a los responsables de la política educativa en todo el mundo medir el nivel de conocimientos y habilidades de los estudiantes de su país en comparación con los de otros países, establecer metas políticas de cara a metas cuantificables obtenidas por otros países, y aprender de las políticas y prácticas aplicadas en otros lugares. (OCDE, s.f., p. 8).

Mientras los análisis de resultados de las pruebas comparativas internacionales se han centrado en la relación entre el rendimiento de los estudiantes y factores como la organización del currículum y el gasto nacional en educación (véase Suter, 2000), más atención debería prestarse a los factores que describen cómo los países utilizan la información obtenida a partir de la participación en pruebas internacionales. De hecho, con el fin de que un país "establezca objetivos políticos de cara a metas cuantificables obtenidas por otros países, y aprender de las políticas y prácticas aplicadas en otros lugares" (ver la cita anterior) de forma adecuada, es necesario hacer interpretaciones de las calificaciones de las pruebas que sean sensibles a los contextos nacionales. El motivo es simple: Las políticas y prácticas que parecen exitosas (como lo reflejan los resultados de PISA) en ciertos países pueden no ser exitosas o pueden ser difíciles de implementar en otros países.

La influencia de PISA en las políticas educativas varía considerablemente de un país a otro (Breakspear, 2012) y puede ser determinada por interpretaciones erróneas y el uso indebido de las puntuaciones y los rankings de los países (Ercikan, Roth, & Asil, 2015). Por otra parte, el impacto de los resultados de PISA en las políticas nacionales no garantiza necesariamente un impacto correspondiente en las prácticas pedagógicas y evaluativas de los países (Teltemann, & Klieme, 2016). Claramente, está emergiendo un panorama complejo de los factores culturales, económicos y sociales que definen la participación en PISA. Comprender la interacción de estos factores es crucial para alcanzar los objetivos de PISA y asegurar que los países se beneficien verdaderamente de su participación en esta prueba comparativa internacional.

Este artículo aborda una importante consideración que subyace al comienzo de las comparaciones de las pruebas internacionales, en los años 1960 – que ciertos recursos humanos e institucionales son necesarios para poder participar de manera apropiada (ver Husén, 1983). Nuestro objetivo es entender la relación entre la capacidad evaluativa nacional y cómo los países participan en las pruebas comparativas internacionales. La capacidad evaluativa es especialmente importante en los países de América Latina, ya que muchos de ellos han empezado a desarrollar programas nacionales de evaluación recientemente, en muchos casos teniendo solo una escasa experiencia en programas de evaluación a gran escala (Ferrer, 2006; Ravela, 2001).

En primer lugar, discutimos cómo la capacidad evaluativa de un país influye en gran medida en el grado en que éste puede beneficiarse por participar en PISA y hacer interpretaciones adecuadas de los resultados de PISA. A continuación, proponemos un marco conceptual para examinar la capacidad evaluativa como clave para abordar dos aspectos de la validez –cultural y

consecuencial. El primero se refiere a la medida en que, durante todo el proceso de desarrollo de la evaluación, se toma en consideración que la experiencia cultural influye en la manera en que los estudiantes interpretan los ítems de una prueba (Solano-Flores, 2011; Solano-Flores y Nelson-Barbero, 2001); el último se refiere a la utilización y a las consecuencias de las inferencias basadas en las calificaciones de una prueba (Messick, 1989; Shepard, 1997). Finalmente, discutimos las múltiples facetas de la capacidad evaluativa como condiciones para abordar la validez cultural y la validez consecuencial en las pruebas comparativas internacionales.

Capacidad Evaluativa

El Programa de las Naciones Unidas para el Desarrollo define *capacidad* como "la facultad de las personas, instituciones y sociedades para desempeñar funciones, resolver problemas y establecer y alcanzar objetivos de una manera sostenible" (Capacity Development Group, 2007, p. 3). Adoptamos y ampliamos los tres aspectos de la capacidad evaluativa identificados por Clarke (2012) en el contexto de las pruebas comparativas:

1. un contexto propicio que apoye o facilite actividades evaluativas – el grado en el que un país ha desarrollado o es capaz de desarrollar y utilizar instrumentos de evaluación técnicamente apropiados;
2. la alineación de actividades e instrumentos evaluativos con otros componentes del sistema educativo – el grado en el que un país ha creado o es capaz de crear y mantener un sistema evaluativo, y
3. la calidad psicométrica de los instrumentos generados – el grado en el que un país es capaz de utilizar la información obtenida mediante aquellos instrumentos y sistemas de evaluación para informar sus políticas y prácticas.

La OECD reconoce la importancia de que los países desarrollen su capacidad evaluativa para poder evaluar eficazmente el aprendizaje de los estudiantes y la necesidad de que analicen sistemáticamente su capacidad para participar en las comparaciones internacionales de las pruebas, ya que "muchos países pueden no tener una buena comprensión de las actividades evaluativas o del tipo gestión involucrado" (Lockheed, Prokic-Bruer, & Shadrova, 2015, p. 60). De acuerdo con estos razonamientos, recursos financieros limitados, una historia corta de participación en evaluaciones de gran escala, una cultura de evaluación incipiente, y un acceso limitado a especialistas con entrenamiento formal en el campo de la psicometría, son factores que debilitan la capacidad evaluativa de un país (Ercikan & Solano-Flores, 2016; Solano-Flores, 2008).

Sostenemos que una capacidad evaluativa limitada dificulta que un país pueda implementar apropiadamente los procedimientos de los programas de pruebas comparativas internacionales y puede impedir que obtenga el máximo beneficio de su participación. Un indicador del posible efecto de una limitada capacidad evaluativa es el hecho de que algunos países participantes en tales programas no cuentan con un número suficiente de expertos nacionales (Kamens & McNeely, 2010). Esta disponibilidad limitada de expertos puede ser una seria dificultad aún en casos en que los profesionales nacionales involucrados en evaluaciones internacionales también participan en evaluaciones nacionales (Gilmore, 2005).

No hay una manera sencilla de determinar el número de expertos en evaluación en un país, si consideramos uno de los múltiples aspectos de la capacidad evaluativa. Sin embargo, con base en la información disponible de PISA 2012 y el directorio de la International Test Commission (ITC) (Illescu, comunicación personal, Noviembre 11, 2012) es posible notar que existe una desigualdad tremenda en la distribución de expertos entre los países participantes en pruebas comparativas internacionales. Quince países que participaron en PISA en 2012 (el 24 por ciento), no tenían, en ese año, a algún individuo con membrecía en la ITC. Estos

números sugieren que un número considerable de países participantes carecen del número adecuado de expertos en evaluación.

Marco Conceptual sobre Capacidad Evaluativa y Validez

La Figura 1 representa nuestro marco conceptual que se establece la relación entre capacidad evaluativa y validez en evaluaciones comparativas internacionales. Para los propósitos de este artículo, se puede pensar que la participación de un país en una prueba comparativa internacional tiene cuatro etapas:

Etapa 1: *Desarrollo de la Prueba*. Los países participantes desarrollan los ítems de las pruebas y los seleccionan para su inclusión de acuerdo con los criterios de formato y contenido especificados por la agencia organizadora correspondiente y con documentos como el marco conceptual de la prueba y las especificaciones de ítems.

Etapa 2: *Traducción de la Prueba*. Los ítems de la prueba son traducidos y adaptados de acuerdo con los lineamientos de traducción proporcionados por la agencia organizadora con el propósito de reflejar las características de la cultura y el tipo de lenguaje utilizado en los currículums nacionales.

Etapa 3: Aplicación de la Prueba y Análisis de Resultados. Las pruebas se aplican a los estudiantes y la agencia organizadora colecta, analiza, y reporta los datos.

Etapa 4: Uso de los Datos de la Evaluación. Los países participantes implementan nuevas políticas educativas basándose en los resultados de los resultados nacionales, a menudo en comparación con el desempeño de otros países, y presumiblemente con el fin de identificar cómo mejorar sus políticas y prácticas educativas.

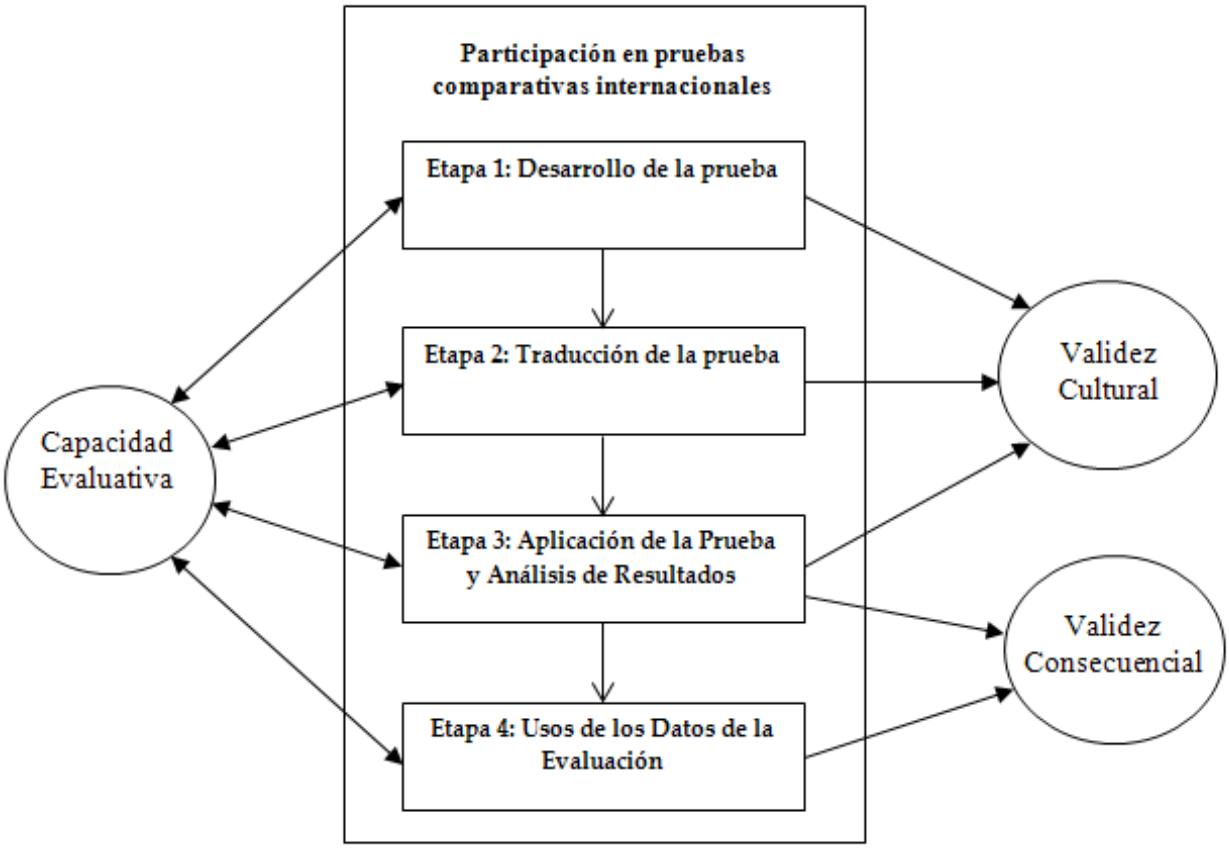


Figura 1. Marco conceptual de la relación entre la capacidad evaluativa nacional y la participación en pruebas comparativas internacionales

Las flechas dobles en la Figura 1 representan la relación sinérgica entre la capacidad nacional de evaluación de un país y su participación en las pruebas comparativas internacionales. Por ejemplo, la capacidad evaluativa influye en la fidelidad con que un país implementa los procedimientos de PISA. A su vez, la participación en PISA contribuye a que el país desarrolle su capacidad evaluativa.

Las flechas sencillas representan el impacto que tiene en la validez la sinergia entre la capacidad evaluativa y las actividades realizadas en las cuatro etapas listadas arriba. En las etapas 1 y 2 (respectivamente, Desarrollo de la Prueba y Traducción de la Prueba), esta sinergia tiene que ver principalmente con la validez cultural. En las etapas 3 y 4 (respectivamente Aplicación de la Prueba y Análisis de Resultados y Uso de los Datos de la Evaluación), esta sinergia tiene que ver principalmente con la validez consecuencial. En la etapa 3, esta sinergia tiene que ver también con la validez cultural.

Validez Cultural

De acuerdo con nuestro marco conceptual, la participación exitosa en las etapas Desarrollo de la Prueba y Traducción de la Prueba contribuye a la validez cultural en el desarrollo y la interpretación de los resultados de las pruebas que se aplican en el país. Para los propósitos de este artículo, podemos definir a la cultura como la historia compartida y el conjunto de experiencias, prácticas, valores, y visiones compartidas en un grupo social y que son mediadas por normas explícitas o implícitas de comunicación y socialización. Esta definición amplia abarca múltiples aspectos de la vida de una sociedad, tales como sus circunstancias económicas o ambientales, idiomas, dialectos, tradiciones, legislación, política, estilos de aprendizaje, prácticas de enseñanza, epistemologías, y formas de representar información (entre muchos otros aspectos) (véase Solano-Flores y Nelson-Barber, 2001).

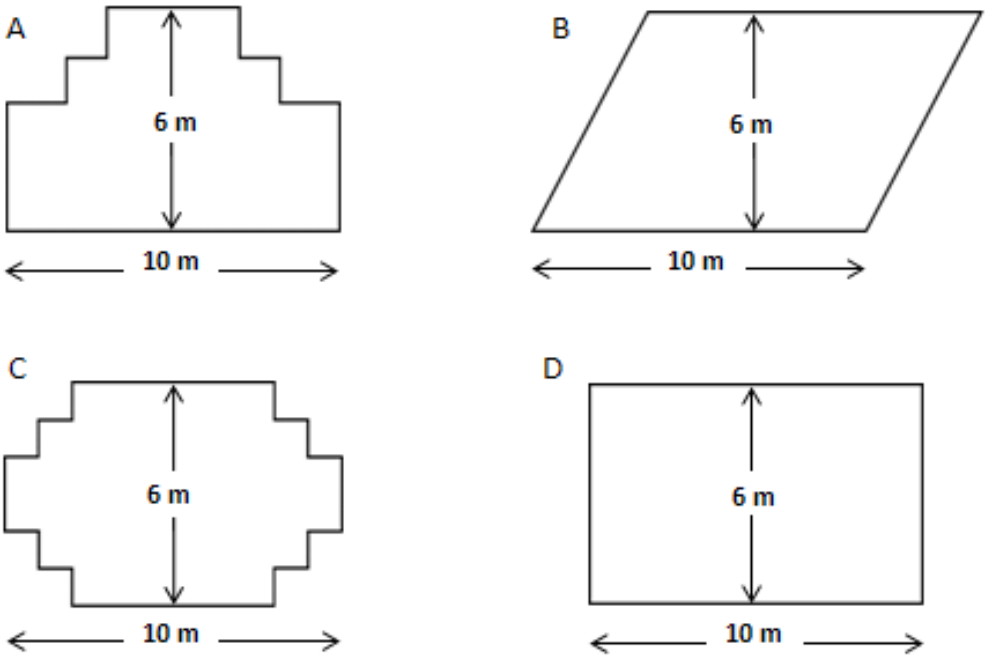
Las teorías contemporáneas en antropología cultural y adquisición del lenguaje sostienen que la cultura y el idioma influyen significativamente en la manera en que la gente hace sentido de sus experiencias y en la manera en que construye significado (p. ej. Bialystok, 2002; Vygotsky, 1978; Wertsch, 1985). Como artefactos culturales, las pruebas no son la excepción: la cultura permea todos los aspectos de la evaluación (Basterra, 2011; Bonnet, 2002; Dogan & Circi, 2010; Hamano, 2011; Solano-Flores, Contreras-Niño, & Backhoff, 2006; Wuttke, 2007). El contenido que los ítems de una prueba pretenden evaluar se basa en formulaciones explícitas sobre lo que los estudiantes de un grado escolar determinado deben saber—formulaciones hechas, por ejemplo, en el marco conceptual de la prueba PISA. En cambio, el formato de esos ítems y la información contextual que proporcional puede estar en gran medida basados en suposiciones implícitas acerca de la experiencia cultural de los examinados.

El ítem que se muestra en la Figura 2 ilustra lo anterior. Mientras que el formato en que los estudiantes dan sus respuestas (esto es, encerrando en un círculo el “Sí” o el “No” para cada categoría) puede estar formalizado en el documento de especificación de ítems, el grado de familiaridad de los estudiantes con este formato puede variar de país a país, dependiendo de las características del currículum, las prácticas pedagógicas formales e informales, y las formas en que las sociedades en que los estudiantes viven representan información. Esto no implica que los estudiantes serían incapaces de entender y responder a los ítems proporcionados en este formato si no están familiarizados con él. Sin embargo, hay evidencia de que estudiantes de diferentes grupos culturales hacen sentido de los ítems de una prueba basándose en distintos de experiencias culturales que tienen lugar fuera de la escuela (Solano-Flores & Li, 2009).

Desde una perspectiva cognitiva, la falta de familiaridad con el formato de una tarea puede incrementar la carga cognitiva—el esfuerzo hecho por la memoria de trabajo (Sweller, 1994). Potencialmente, los estudiantes que están familiarizados con el formato pueden necesitar usar más memoria de trabajo que los estudiantes que están familiarizados con el formato para poder hacer sentido de la tarea que tienen que completar y decidir cómo dar sus respuestas. Esta diferencia en el tiempo necesario para procesar información puede ser una fuente de sesgo.

Pregunta 1: EL CARPINTERO

Un carpintero tiene 32 m de madera y quiere hacer una cerca alrededor de una jardinera. Está considerando los siguientes diseños para la jardinera.



Para cada diseño, encierra en un círculo el "Sí" o el "No" para indicar si se puede hacer la cerca con 32 metros de madera.

Diseño de la jardinera	Usando este diseño ¿se puede construir la cerca con 32 metros de madera?
Diseño A	Sí / No
Diseño B	Sí / No
Diseño C	Sí / No
Diseño D	Sí / No

Figura 2. Ítem del carpintero. Fuente: OECD (2006)

Por supuesto, los procedimientos de las agencias que organizan las pruebas internacionales toman en cuenta estas diferencias. Por ejemplo, la creación de ítems es un proceso colaborativo en el que los representantes nacionales interactúan con

comités asesores expertos, tienen la oportunidad de tomar acciones relacionadas con la validez cultural de la prueba, y pueden, hasta cierto grado, hacer ciertas adaptaciones a los contextos culturales nacionales (p. ej. OECD, 2010). De hecho, se recomienda a los países "poner especial atención a aspectos relaconados con nacionalidad, cultura, identidad, etenicidad y locación geográfica" (Mullis & Martin, 2011, p. 8). Sin embargo, a pesar de estos esfuerzos, hay aún mucho por aprender sobre validez cultural en pruebas comparativas internacionales. Por ejemplo, típicamente, los equipos de expertos a cargo de revisar ítems (p. ej. Mullis, Martin, Ruddo, O'Sullivan & Preuschoff, 2009) no incluyen expertos en disciplinas relacionadas con el lenguaje y la cultura. Sin embargo, existe evidencia en el campo de la evaluación de minnorías lingüísticas de que la participación de individuos de diferentes culturas en el proceso de desarrollo de una prueba, aunque necesaria, no garantiza que se cubran adecuadamente los aspectos culturales de una prueba (Solano-Flores y Gustafson, 2013). En la ausencia de un entrenamiento formal y especializado, tales profesionales pueden no estar totalmente conscientes de las formas sutiles en que la cultura influye en las interpretaciones que hacen los estudiantes de los ítems, o del grado en que la falta de familiaridad con la información contextual que proporcionan los ítems puede dificultar a los estudiantes hacer sentido de los ítems.

De acuerdo con nuestro marco conceptual, llevar exitosamente a cabo las actividades que forman parte de la etapa, Aplicación de la Prueba y Análisis de Resultados contribuye también a la validez cultural. Una de estas actividades es el uso de la teoría de la respuesta al ítem en la detección de ítems que tienen sesgo cultural. Un ítem tiene sesgo (o funciona diferencialmente) si dos grupos diferentes, el grupo de referencia y el grupo focal (p. ej., respectivamente el grupo al que se evalúa con la version original de una prueba y el grupo al que se evalúa con la traducción de esa prueba) tienen diferentes probabilidades de responder correctamente a ese ítem una vez que se controlan las diferencias que los grupos tienen en su desempeño global en la prueba (Camilli, 2006; Camilli & Shepard, 1994). A pesar de que, al discutir sesgo en contextos multiculturales, frecuentemente se menciona al funcionamiento diferencial de ítems (p. ej. van de Vijver, 2016), su efectivdad está limitada por el grado en que la técnica se usa con números substanciales de ítems y en etapas en el desarrollo de pruebas en que los ítems con sesgo pueden ser eliminados o modificados oportunamente. Obviamente, una capacidad evaluativa limitada puede impedir que los países usen esta técnica apropiadamente o de manera suficiente.

Validez Consecuencial

Una capacidad evaluativa limitada puede afectar al grado con que un país saca provecho de la información obtenida a partir de las actividades de evaluación. Aunque muchos países usan los resultados de pruebas internacionales para sustentar sus reformas educativas (Gilmore, 2005; Stachelek, 2010), la mayoría usa principalmente los rankings como evidencia de que la educación ha fallado o ha tenido éxito. Varios estudiosos han advertido que los rankings deben ser usados con cuidado y que se les debe usar solamente cuando se entiende a fondo cómo se construyen (Figazzolo, 2009; Hamano, 2008/2011; Sjøberg, 2007; Stachelek, 2010; Tatto, 2006; Wuttke, 2007). Desgraciadamente, el uso estratégico y adecuado de esta información parece improbable en circunstancias en que la capacidad evaluativa es limitada.

De acuerdo con nuestro marco conceptual, la participación exitosa en las etapas, Aplicación de la Prueba y Análisis de Resultados y Uso de los Datos de la Evaluación (Figura 1) contribuye a la validez consecuencial. Messick (1989) definió a la validez en general como "un juicio evaluativo integrado del grado en que la evidencia empírica y los razonamientos teóricos apoyan la adecuación y *propiedad* de las *inferencias* y *acciones* basadas en las calificaciones en

pruebas y en otras formas de evaluación" (p. 13, curisvas en el original). Shepard (1997) incluye el uso y las consecuencias de las inferencias basadas en calificaciones de pruebas como parte del concepto de validez consecuencial.

Aunque actualmente se debate si las acciones basadas en las calificaciones en pruebas debieran formar parte del concepto de validez, Messick (1995) hizo énfasis en el aspecto social del uso de las pruebas. Él argumenta que "para evaluar qué tan bien una prueba cumple su función, uno debe preguntarse no solamente si las consecuencias sociales potenciales y reales de la interpretación de la prueba y su uso apoyan los propósitos originales de la prueba sino también si esas consecuencias son consistentes con otros valores sociales" (p.744).

Estas importantes nociones, que son críticas para la evaluación válida y justa de diversos grupos culturales, no reciben suficiente atención en el contexto de las pruebas nacionales (véase Kane, 2006) y no han sido examinadas en el contexto de las pruebas comparativas internacionales. Algunos de los muchos aspectos que deben ser examinados son: *¿En qué medida influye la capacidad evaluativa en la fidelidad con que un país implementa procedimientos de programas evaluativos internacionales? ¿De qué manera la participación en una prueba comparativa internacional contribuye a desarrollar la capacidad evaluativa de un país? ¿Cómo pueden asegurarse las autoridades educativas y quienes deciden las políticas educativas de que las inferencias que hacen de los resultados de una prueba son válidas y de que las reformas que establecen no tienen consecuencias negativas imprevistas?*

Desafortunadamente, aunque al parecer existe la suposición implícita de que los países participantes en PISA tienen sistemas evaluativos adecuados, no existen evaluaciones que hayan examinado cómo los países desarrollan su capacidad evaluativa a partir de su participación en PISA (Lockheed et al., 2015).

Un aspecto importante de la relación entre validez consecuencial y capacidad evaluativa es que las respuestas a preguntas como las formuladas arriba y la forma en que se atienden diversos problemas relacionados con la evaluación varían de país a país. Hay evidencia sólida de que el significado de lo que es una prueba y las consecuencias y posibilidades asociadas a las pruebas son muy diferentes en distintos países. El rango es amplio: desde la percepción de las pruebas como instrumentos de opresión hasta la percepción de las pruebas como una oportunidad para la promoción social (e.g., see Gebril, 2016; Kennedy, 2016; Lingard & Lewis, 2016).

Condiciones para la Validez Cultural y la Validez Consecuencial

Basándonos en el razonamiento expuesto, es posible identificar las principales condiciones que contribuyen a la validez cultural y consecuencial en las pruebas comparativas internacionales. La Tabla 1 proporciona una lista inicial, con adaptaciones, de los componentes que han sido desarrollados en diversos esfuerzos anteriores para examinar validez cultural, validez consecuencial, y la capacidad evaluativa en contextos nacionales (ver Ad-Hoc Technical Committee on the Development of Technical Criteria for Examining Cultural Validity in Educational Assessment, 2015; Martínez-Rizo, 2015) e internacionales (Ercikan & Solano-Flores, 2016; Solano-Flores, 2008). Las condiciones se agrupan en tres categorías: Programa de Pruebas Comparativas Internacionales, Participación y Práctica.

Tabla 1. *Condiciones para la Validez Cultural y la Validez Consecuencial*

PROGRAMA DE PRUEBAS COMPARATIVAS INTERNACIONALES
Marco de Conceptual de la Prueba. El desarrollo del marco de la prueba toma en cuenta que, siendo ésta un artefacto cultural, el desempeño de los estudiantes está mediado por su experiencia sociocultural y lingüística.
Especificación de la Muestra Poblacional. Los procedimientos utilizados para definir y extraer muestras de la población de estudiantes son sensibles al hecho de que los países tienen distintas formas de diversidad cultural y lingüística y diferentes contextos sociales y escolares.
Especificaciones de Ítems. Además de estar basadas en las características de las competencias y los conocimientos evaluados, las decisions relativas a los diferentes formatos de ítems utilizados en la prueba toman en cuenta la noción de que los estudiantes de diferentes culturas pueden no estar igualmente familiarizados con la información contextual proporcionada por los ítems y sus elementos lingüísticos y gráficos.
Mecanismos de Corrección. El proceso de desarrollo de la prueba estipula las acciones que deben ser tomadas cuando se detectan ítems inadecuados de acuerdo con las evidencias de validez cultural y cognitiva obtenidas a partir de diferentes fuentes, incluyendo la revisión de expertos, el análisis del funcionamiento diferencial de los ítems y los estudios de generalizabilidad.
PARTICIPACIÓN
Agenda de Investigación y Práctica: El país es capaz de utilizar los resultados de su participación en la prueba para generar nuevos conocimientos sobre temas de interés nacional y para mejorar la práctica en esas áreas.
Sostenibilidad, Estabilidad y Continuidad: El país es capaz de sostener programas a largo plazo y actividades relacionadas con su participación en el programa de evaluación, independientemente de cualquier incertidumbre financiera o política.
Recursos Humanos: El país tiene o es capaz de desarrollar o aumentar en un tiempo razonable una masa crítica de profesionales calificados en el campo de la medición educativa y áreas relacionadas en relación con su participación en el programa de evaluación.
Recursos Económicos e Infraestructura: El país cuenta con la infraestructura mínima y es capaz de asignar los recursos económicos necesarios para llevar a cabo actividades relacionadas con su participación en la prueba después de que su participación en la misma ha terminado.
Congruencia Sistémica: El país coordina su participación en la prueba con componentes clave de su sistema educativo.
Toma de Decisiones: El país dedica recursos y tiempo suficientes para hacer una interpretación cuidadosa de los resultados de la evaluación (por ejemplo, más allá de los rankings de los países) y toma decisiones adecuadas en relación con su sistema educativo dentro del alcance de la información que la evaluación es capaz de producir.
Implementación y Enriquecimiento: Además de implementar con fidelidad los procedimientos establecidos por el programa de evaluación, el país es capaz de añadir actividades a aquellos procedimientos que ayudan a satisfacer sus necesidades específicas.
PRÁCTICAS DE EVALUACIÓN
Fundamentación Conceptual. Los procedimientos utilizados para tomar en cuenta la diversidad lingüística, cultural y socioeconómica tienen una fundamentación teórica sólida.
Cronograma. El calendario del proceso de desarrollo de la prueba asigna un tiempo razonable para tomar todas las medidas concernientes a la diversidad cultural.
Equipo de Desarrollo. Además de expertos en el contenido, educadores y diseñadores de pruebas, los equipos encargados de desarrollar los ítems de la prueba incluyen profesionales con especialidades en el campo de la cultura y el idioma (por ejemplo, antropología y scociolinguiística), así como representantes de diversos grupos culturales y lingüísticos.
Representación de la Diversidad en la Muestra de Población. Las muestras de estudiantes incluyen, en

todas las etapas del proceso del desarrollo de la prueba (por ejemplo, las etapas de pilotaje), diversos grupos culturales, lingüísticos y socioeconómicos.

Entrevistas Cognitivas. Se llevan a cabo entrevistas cognitivas para examinar si los estudiantes de diferentes grupos culturales, lingüísticos y socioeconómicos interpretan los ítems de la misma manera.

Proceso de Revisión por Expertos. Existe un proceso de revisión durante el cual equipos de expertos externos examinan los ítems y posibles demandas irrelevantes a los constructos para estudiantes de diferentes grupos culturales, lingüísticos y socioeconómicos.

Análisis del Funcionamiento Diferencial de los Ítems. Se examinan muestras representativas de los ítems para determinar si tienen un funcionamiento diferencial en detrimento de grupos focales especiales (por ejemplo, estudiantes de bajos ingresos y estudiantes de grupos culturales y lingüísticos específicos).

Estudios de Generalizabilidad. Se lleva a cabo una serie de estudios de generalizabilidad con diferentes combinaciones de ítems y diferentes muestras de la población estudiantil para examinar la comparabilidad de los coeficientes de generalizabilidad (fiabilidad) obtenidos con distintos grupos culturales, lingüísticos y socioeconómicos.

Desagregación de Datos. Las características técnicas de la prueba se examinan por separado para cada grupo cultural, lingüístico y socioeconómico importante.

Programa de Pruebas Comparativas Internacionales

Las condiciones en la categoría Programa de Pruebas Comparativas Internacionales se refieren a actividades, procedimientos o productos ya existentes, al menos en principio, pero que deben abordar los aspectos de la cultura y diversidad de manera más explícita. Sostenemos que el tratamiento de la validez cultural depende en gran medida del grado en el que se considera a la cultura en el proceso de elaboración de documentos normativos y por medio de la planificación de las actividades del programa de evaluación.

Los marcos conceptuales de pruebas internacionales existentes no son ingenuos al hecho de que la cultura puede influir en la manera en que los estudiantes interpretan los ítems y generalmente contienen una discusión sobre aspectos culturales. Sin embargo, un marco conceptual podría hacer una contribución más sustancial si tratara a la cultura como un componente integrante del conocimiento, no algo en lo que se piensa después de haber desarrollado la mayor parte del documento.

Otras condiciones del Programa de Evaluación que podrían parecer triviales pueden en realidad ser críticas para atender adecuadamente a la validez cultural. Tal es el caso de *Cronograma* y *Mecanismos de Corrección*. Por lo general, las acciones destinadas a abordar las cuestiones relacionadas con cultura (p. ej. la revisión de la traducción o los procedimientos para examinar la sensibilidad cultural o identificar fuentes potenciales de sesgo) son evaluadas de forma sumativa (al final del proceso del desarrollo de evaluación), no formativa (durante el proceso del desarrollo de evaluación). Una consecuencia potencial indeseable de esta práctica es que los retrasos en el completamiento de diferentes etapas del proceso de desarrollo de la prueba se acumulen en detrimento del tiempo asignado para abordar cuestiones culturales.

Participación

Las condiciones de la categoría Participación se refieren a las características de aquellas acciones adoptadas por el país más allá de la simple implementación de los procedimientos del programa de evaluación, y a la capacidad del país para insertar estas acciones como parte de un plan nacional más amplio. Sostenemos que abordar tanto la validez cultural como la validez consecuencial depende en gran medida del grado en que la participación de un país es impulsada por una idea clara de lo que hay que mejorar y lo que se necesita, económicamente y en términos de

recursos humanos, para hacer cambios necesarios. Ser capaz de dar sentido a los datos de una evaluación internacional, más allá de las conclusiones basadas en los rankings de los países, requiere que un país haga las asignaciones adecuadas de tiempo y de recursos humanos, materiales y económicos. El uso adecuado de la información que se obtiene en la aplicación de una prueba comparativa internacional, independientemente de cualquier presión política, requiere de un mínimo de estabilidad institucional y de una congruencia del sistema educativo nacional. Puede requerir también de transformaciones políticas profundas, aquellas que tienen que ver con la estructura de la asignación de fondos en apoyo a las escuelas con altas necesidades (Darling-Hammond, 2014). El uso adecuado de la información obtenida mediante la aplicación de una prueba comparativa internacional también implica una interpretación de los resultados como reflejo de las profundas desigualdades sociales (Carnoy, 2015).

Si prestamos atención especial a *Implementación y Enriquecimiento*, podemos comprender el grado en el que los países pueden maximizar el beneficio de participar en pruebas comparativas internacionales. Los ítems de la prueba PISA se crean inicialmente en inglés y en francés, y después son traducidos de esos dos idiomas fuente a los idiomas de los países participantes. A los países participantes en PISA se les proporcionan lineamientos para la traducción de las pruebas (p. ej. Hambleton, 2005; National Project Managers' Meeting, 2010) y existen mecanismos bien establecidos para revisar la adecuación de la traducción de los ítems. Sin embargo, hay evidencia de que, debido al hecho de que las lenguas codifican las experiencias de maneras diferentes, incluso una traducción impecable no puede evitar errores de traducción. El error de traducción debido a la especificidad cultural de la información contextual proporcionada por los ítems y el error debido a la complejidad sintáctica, semántica, y la alteración de los constructos medidos puede aumentar la dificultad del ítem (Solano-Flores, Backhoff, & Contreras-Niño, 2009; 2013). Una implicación de este hallazgo es que, a pesar de que las directrices de traducción proporcionan un estándar valioso y un procedimiento general para la traducción de una prueba, los países necesitan desarrollar estrategias para implementar los procedimientos de traducción y revisión interna de sus propios ítems traducidos que sean sensibles a su cultura y a las variedades específicas de sus idiomas. Tal implementación puede variar enormemente entre los países en función de sus recursos y de su capacidad para reunir equipos adecuados de traducción de las pruebas.

Prácticas

Las condiciones en la categoría Prácticas se refieren a las actividades evaluativas que un país es capaz de realizar no sólo cuando participa en una prueba comparativa internacional, sino también como parte de sus propios programas de evaluación. Sostenemos que abordar cuestiones relacionadas con la validez cultural y con la validez consecuencial depende en gran medida del grado en que las prácticas y las experiencia que tienen los países en el área de evaluación son sensibles a aspectos de la cultura en sus propios contextos nacionales.

Al igual que las condiciones enumeradas en la categoría *Programas de Evaluación*, la mayoría de las condiciones enumeradas en esta categoría ya existen en principio. Por ejemplo, como se mencionó anteriormente, el sesgo cultural se puede examinar mediante técnicas para la detección del funcionamiento diferencial de ítems (p. ej. van de Vijver, 2016). Sin embargo, las posibilidades de estas técnicas no garantizan su uso. No es claro en qué medida los países participantes en PISA examinan el funcionamiento diferencial de los ítems de forma rutinaria con un número sustancial de ítems o qué acciones toman cuando identifican ítems con sesgo. Por otra parte, si bien es necesario, el análisis del funcionamiento diferencial de ítems puede no detectar ítems con sesgo cuando la

heterogeneidad cultural y lingüística en los grupos focales no se toma en consideración. Existe evidencia de que cuanto más heterogéneo lingüísticamente y culturalmente es el grupo focal, es menos probable identificar ítems que funcionan diferencialmente (Ercikan, Roth, Simon, Sandilands & Lyons-Thomas, 2014). Una limitada capacidad evaluativa puede obstaculizar la capacidad de algunos países para examinar el funcionamiento diferencial de ítems para grupos focales importantes y para modelar adecuadamente la heterogeneidad de sus poblaciones.

La condición, *Estudios de Generalizabilidad* merece especial atención. Actualmente los estudios de generalizabilidad no forman parte de la práctica establecida para examinar aspectos de cultura, aunque la investigación muestra su importancia en el análisis de la validez cultural (Solano-Flores & Li, 2006, 2009). La teoría de la generalizabilidad (G) (Brennan, 2001; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991) es una teoría del error de medición y también del muestreo de observaciones (Kane, 1982). A diferencia de los enfoques basados en el análisis del funcionamiento diferencial de ítems, los enfoques basados en la teoría G se basan en el análisis de las calificaciones globales de las pruebas asumiendo la composición aleatoria de la prueba (ver Solano-Flores, 2016). La investigación sobre el uso de la teoría G en la evaluación de minorías lingüísticas y culturales indica que la fiabilidad de las puntuaciones en la misma prueba puede ser diferente para distintos grupos culturales. La implicación de este hallazgo lleva a otra condición – *Desagregación de los Datos*-. Desagregar los datos por grupos culturales y comparar los grupos en términos de la calidad psicométrica de sus puntuaciones en la prueba es una estrategia más rigurosa que simplemente comparar grupos en términos de las medias de sus calificaciones (Solano-Flores & Li, 2013).

Observaciones Finales

Hemos examinado la relación entre la capacidad evaluativa, la validez cultural y la validez consecuencial en PISA. Sostenemos que un nivel crítico de la capacidad evaluativa es necesario para que un país se beneficie de participar en pruebas comparativas internacionales. Al mismo tiempo, la participación en las pruebas internacionales es una buena oportunidad para que los países aumenten su capacidad evaluativa, siempre que presten debida atención a las cuestiones de validez cultural y validez consecuencial.

Algunas consideraciones sobre la validez cultural y consecuencial deberían influir en los procedimientos utilizados por las agencias internacionales que organizan las pruebas comparativas internacionales (p. ej. aquellas concernientes al desarrollo del marco conceptual de una prueba y los documentos de especificaciones de los ítems). Otras implican la preparación o la responsabilidad de cada país participante para hacer frente a los desafíos que son específicos a la composición cultural de su población estudiantil.

Una capacidad evaluativa limitada puede dificultar la efectividad con la que un país da atención adecuada a la validez cultural y sin duda afecta a la validez consecuencial. Esto es una cuestión fundamental de equidad a nivel internacional. Los países participantes en PISA necesitan asegurar que tienen una capacidad evaluativa mínima, si su participación en las pruebas comparativas internacionales ha de informar con precisión sus esfuerzos para efectuar reformas educativas. De no ser el caso, la participación de los países participantes en PISA tiene que estar acompañada por programas sólidos para el desarrollo de la capacidad evaluativa nacional. Lo más importante: las agencias organizadoras deben asumir la responsabilidad de apoyar a los países en el desarrollo de su capacidad evaluativa.

Referencias

Ad-Hoc Technical Committee on the Development of Technical Criteria for Examining Cultural Validity in Educational Assessment. (2015). Promoting and evaluating cultural validity in the activities performed by the National Institute for Educational Evaluation (INEE). Submitted to the National Institute for Educational Evaluation. Mexico City, Mexico, January 16.

Basterra, M. R. (2011). Cognition, culture, language, and assessment. En M. R. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 72-95). New York: Routledge.

Bialystok, E. (2002). Cognitive processes of L2 users. En V. J. Cook (Ed.), *Portraits of the L2 user* (pp. 145-165). Buffalo, NY: Multilingual Matters.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice*, *9*(3), 387-399.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Paper Number 71*. Consultado en OECD website: http://www.oecd-ilibrary.org/education/the-policy-impact-of-pisa_5k9fdfqffr28-en

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer Verlag.

Camilli, G. (2006). Test fairness. En R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). Westport, CT: American Council on Education and Praeger.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousands Oaks, CA: Sage.

Capacity Development Group (2007, May). *Capacity assessment methodology: User's guide*. Bureau for Development Policy, United Nations Development Program. New York, September 2005. Consultado en the United Nations Development Programme website: https://www.unpei.org/sites/default/files/PDF/institutioncapacity /UNDP-Capacity-Assessment-User-Guide.pdf

Carnoy, M. (2015). International test score comparisons and educational policy. Carnoy, M. (2015). *International Test Score Comparisons and Educational Policy: A Review of the Critiques*. Boulder, CO: National Education Policy Center. Consultado en http://nepc.colorado.edu/publication/international-test-scores

Clarke, M. (2012). What matters most for student assessment systems: A framework paper. Consultado en the World Bank website: https://openknowledge.worldbank.org/bitstream/handle/10986/17471/682350WP00PUBL0WP10READ0web04019012.pdf?sequence=1

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Darling-Hammond, Linda (2014). What can PISA tell us about U.S. education policy? *New England Journal of Public Policy*: *26*(1), Art. 4. Consultado en http://scholarworks.umb.edu/nejpp/vol26/iss1/4

Dogan, E., & Circi, R. (2010). A blind item-review process as a method to investigate invalid moderators of item difficulty in translated assessment. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 3) (pp. 157-172). Hamburg: IERI.

Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about uses of international assessments. *Teachers College Record, 117*(1), 1-28.

Ercikan, K., Roth, W.-M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education, 27*, 275-285.

Ercikan, K., & Solano-Flores, G. (2016). Assessment and sociocultural context: A bidirectional relationship. En G. T. L. Brown & L. Harris (Eds.), *Human Factors and Social Conditions of Assessment*. New York: Routledge.

Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges*. Washington, DC: PREAL. Consultado en http://www.uis.unesco.org/Education/Documents/Ferrer.pdf

Figazzolo, L. (2009). *Impact of PISA 2006 on the education policy debate*. Consultado en http://download.ei-ie.org/docs/IRISDocuments/Research%20Website%20Documents/2009-00036-01-E.pdf

Gebril, A. (2016). Educational assessment in Muslim countries: Values, policies, and practices. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Gilmore, A. (2005). The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: An evaluation of the value of World Back support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS). Consultado en http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/Gilmore_Impact_PIRLS_TIMSS.pdf

Hamano, T. (2011).The globalization of student assessments and its impact on education policy [English version]. *Proceedings, 13*, 1-11. (Originalmente apareció en japonés en 2008 en el *Annual Bulletin of JASEP (Japan Academic Society for Educational Policy), 15*, 21-37). Consultado en http://teapot.lib.ocha.ac.jp/ocha/bitstream/10083/51418/1/Proceedings13_01Hamano.pdf

Hambleton, R.K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. En R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.

Husén, T. (1983). *An incurable academic: Memoirs of a professor*. Oxford, UK: Pergamon Press.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5-25. doi: http://dx.doi.org/10.1086/648471

Kane, M. T. (1982). A sampling model of validity. *Applied Psychological Measurement, 6*, 125-160. doi: http://dx.doi.org/10.1177/014662168200600201

Kane, M. T. (2006). *Validation*. En R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kennedy, K. J. (2016). Exploring the influence of culture on assessment: The case of teachers' conceptions of assessment in Confucian-heritage societies. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Lingard, B., & Lewis, S. (2016). Globalization of the American approach to accountability: The high price of testing. En G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Martínez-Rizo, F. (2015). Las pruebas ENLACE y EXCALE: Un estudio de validación. Consultado en http://publicaciones.inee.edu.mx/buscadorPub/P1/C/148/P1C148.pdf

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education, Macmillan.

Messick, S. (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as

scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: http://dx.doi.org/10.1037/0003-066X.50.9.741

Mullis, I. V. S., & Martin, M. O. (2011). *TIMSS 2011 item writing guidelines.* Consultado en http://timssandpirls.bc.edu/methods/pdf/T11_Item_writing_guidelines.pdf

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011: Assessment frameworks.* Consultado en http://timssandpirls.bc.edu/timss2011/downloads/TIMSS2011_Frameworks.pdf

National Project Managers' Meeting (2010, October). Translation and adaptation guidelines for PISA 2012. Doc: NPM10104e. PISA Consortium. Budapest, Hungary. Consultado en https://www.oecd.org/pisa/pisaproducts/49273486.pdf

Organisation for Economic Co-operation and Development (OECD). (n.d.). *Programme for international student assessment (PISA): Results from PISA 2012, Country note: United States.* Consultado en http://www.oecd.org/pisa/keyfindings/PISA-2012-results-US.pdf

Organisation for Economic Co-operation and Development (2006). *PISA released items: Mathematics*. Consultado en http://www.oecd.org/pisa/38709418.pdf

Organisation for Economic Co-operation and Development (2010). *Translation and adaptation guidelines for PISA 2012.* Consultado en http://www.oecd.org/pisa/pisaproducts/49273486.pdf

Lockheed, M., Prokic-Bruer, T., & Shadrova, A. (2015). *The experience of middle-income countries participating in PISA 2000-2015* (PISA series). Washington, D.C. & Paris: The World Bank & OECD Publishing. doi: http://dx.doi.org/10.1787 /9789264246195-en

Ravela, P. (Ed.). (2001). Los próximos pasos: ¿Hacia dónde y cómo avanzar en la evaluación de aprendizajes en América Latina? Document No. 20. Working Group on Assessment and Standards. Santiago: PREAL. Consultado en http://campus-oei.org/calidad/grade.PDF

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13. doi: http://dx.doi.org/10.1111/j.1745-3992.1997.tb00585.x

Sjøberg, S. (2007). PISA and "real life challenges": Mission impossible. In S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *According to PISA—Does PISA keep what it promises?* Berlin: LIT Verlag.

Solano-Flores, G. (2008, July). A conceptual framework for examining the assessment capacity of countries in an era of globalization, accountability, and international test comparisons. Comunicación presentada en la *6th Conference of the International Test Commission*, Liverpool, UK.

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. R. Basterra, E. Trumbull, and G. Solano-Flores, *Cultural validity in assessment* (pp. 3-21). New York: Routledge.

Solano-Flores, G. (2016). Generalizability. En L. E. Suter, D. Wyse, E. Smith, & N. Selwyn (Eds.), *The BERA/SSAGE Handbook of Educational Research* (chap. 47). London: Sage.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales. *Revista Electrónica de Investigación Educativa (REDIE), 8*(2). Consultado en http://redie.uabc.mx/redie/article/download/143/246

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L.A. (2009). Theory of test translation

error. *International Journal of Testing, 9*, 78-91.

Solano-Flores, G., Contreras-Niño, L.A., & Backhoff, E. (2013). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. En M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research in the context of the programme for international student assessment* (pp. 71-85). Springer Verlag.

Solano-Flores, G., & Gustafson, M. (2013). Assessment of English language learners: A critical, probabilistic, systemic view. En M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scaleassessment in education: Theory, issues, and practice* (pp. 87-109). New York: Routledge.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, *25*(1), 13-22.

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice*, *28* (2), 9-18.

Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation*, *19*(2-3), 245-263. doi: http://dx.doi.org/10.1080/13803611.2013.767632

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 533-573. doi: http://dx.doi.org/10.1002/tea.1018

Stachelek, A. J. (2010). Exploring motivational factors for educational reform: Do international comparisons dictate educational policy? *Journal of Mathematics Education at Teachers College*, *1*,52-55.

Suter, Larry E. (2000). Is student achievement immutable? Evidence from international studies on schooling and student achievement. *Review of Educational Research*, *70*(4), 529-545. doi: http://dx.doi.org/10.3102/00346543070004529

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*, 295-312. doi: http://dx.doi.org/10.1016/0959-4752(94)90003-5

Tatto, M. T. (2006). Education reform and the global regulation of teachers' education, development and work: A cross-cultural analysis. *International Journal of Educational Research*, *45*, 231-241. doi: http://dx.doi.org/10.1016/j.ijer.2007.02.003

Teltemann, J., & Klieme, E. (2016). The impact of international testing projects on policy and practice. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment* (Chap. 21). New York: Routledge.

van de Vijver, F. J. R. (2016). Assessment in education in multicultural populations. En G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*, (Chap. 25).New York: Routledge.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.

Wuttke, J. (2007). Uncertainties and bias in PISA. En S. T. Hopmann, G. Brinek, and M. Retzl (Eds.), *According to PISA – Does PISA keep what it promises?* Berlin: LIT Verlag.

Author / Autor

To know more / Saber más

Solano-Flores, Guillermo (gsolanof@stanford.edu).

Es profesor en la Escuela de Graduados en Educación de la Universidad de Stanford, Estados Unidos. Es el autor de contacto para este artículo. Se especializa en la evaluación educativa y las cuestiones lingüísticas y culturales que son relevantes tanto para las pruebas comparativas internacionales como para la evaluación de las minorías culturales y lingüísticas. Sus contribuciones al campo de la evaluación educativa incluyen la teoría del error de traducción del test, el uso de la teoría de la generalizabilidad -una teoría psicométrica de error de medida- en las evaluaciones de las minorías lingüísticas, la formalización del concepto de validez cultural, y el desarrollo de una metodología para el diseño y el análisis de las ilustraciones utilizadas en los ítems de los tests. Su dirección postal es: Stanford University Graduate School of Education. 485 Lasuen Mall. Stanford, CA 94305-3096. United.

Milbourn, Tamara (tamara.milbourn@colorado.edu)

Doctoranda en Fundamentos, Política y Práctica de la Educación, de la Universidad de Colorado en Boulder. Tiene una Maestría en Lingüística Aplicada al Inglés por la Universidad de Wisconsin-Madison, Estados Unidos. Su trabajo examina las experiencias de los estudiantes internacionales en las universidades estadounidenses. Se interesa en los problemas de la educación relacionados con el mono / multilingüismo, con énfasis en temas de equidad como los relacionados con las prácticas del lenguaje y las normas académicas. Ha trabajado en Taiwán, Japón y Benin y actualmente enseña cursos de lingüística y política educativa en el sistema de la Universidad de Colorado.

RELIEVE

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

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]