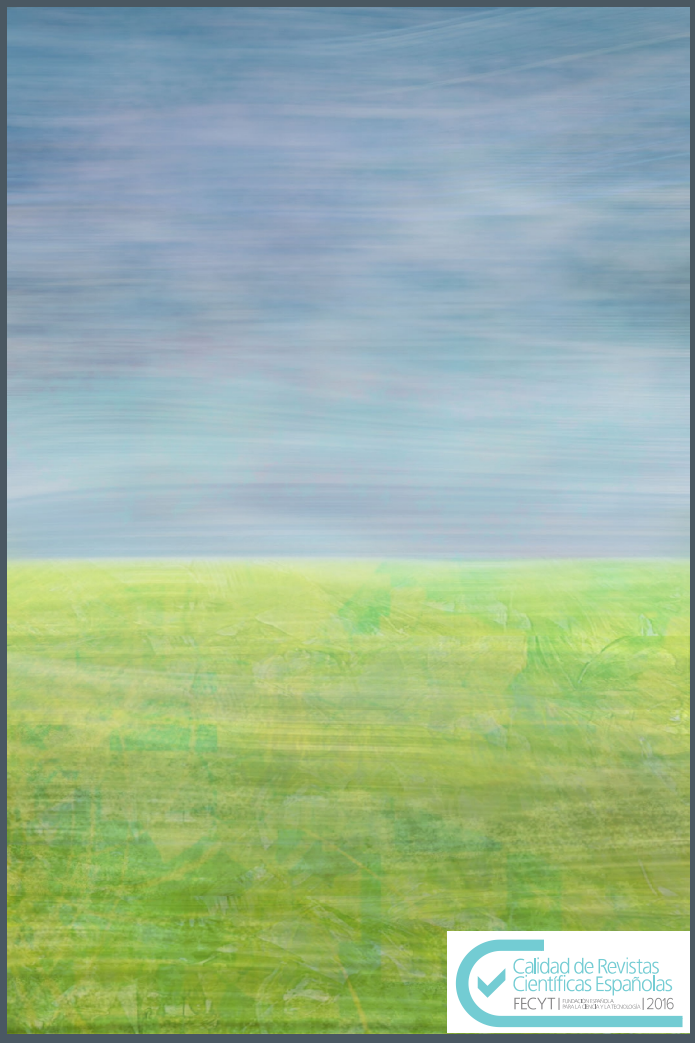
revista de EDUCACIÓN
Nº 383 ENERO-MARZO 2019

Pensémoslo de nuevo: ¿Podemos comparar las escalas de antecedentes socioeconómicos?

Back to the drawing board: Can we compare socioeconomic background scales?

Andrés Sandoval-Hernandez
David Rutkowski
Tyler Matta
Daniel Miranda



GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL

Calidad de Revistas Científicas Españolas
FECYT | 2016

Pensémoslo de nuevo: ¿Podemos comparar las escalas de antecedentes socioeconómicos?[1]

Back to the drawing board: Can we compare socioeconomic background scales?

DOI: 10.4438/1988-592X-RE-2019-383-400

Andrés Sandoval-Hernandez
University of Bath
David Rutkowski
Indiana University
University of Oslo
Tyler Matta
Pearson
University of Oslo
Daniel Miranda
Pontificia Universidad Católica de Chile, MIDE UC
Centro de Estudios de Conflicto y Cohesión Social

Resumen

Utilizando datos de evaluaciones internacionales de gran escala, evaluamos la consistencia y la invarianza de las escalas de antecedentes socioeconómicos de los estudiantes entre los países participantes en estos estudios. Para ello, utilizamos las medidas de antecedentes socioeconómicos desarrolladas por PISA, TERCE y TIMSS, ya que cada estudio operacionaliza esta medida de manera diferente. Como parte de nuestro análisis, examinamos si la escala de TERCE, un estudio latinoamericano –con medidas que fueron desarrolladas con

(1) Este artículo fue apoyado por el Norwegian Research Council Grant #255246 y por el Centro de Estudios de Conflicto y Cohesión Social —COES CONICYT/FONDAP N°°15130009

Recibido: 27-07-2017 Aceptado: 14-09-2018

enfoque regional– exhibe mejores propiedades psicométricas que otras medidas que fueron diseñadas para funcionar en un número mayor y más diverso de sistemas educativos. Adicionalmente, examinamos la escala de TIMSS, un estudio enfocado en tendencias –que históricamente ha enfatizado la comparabilidad y consistencia–. Finalmente, incluimos también la escala de PISA, que cuenta con el mayor número de participantes y que, en cierta medida, ha cambiado y conceptualizado sus cuestionarios de contexto dependiendo el dominio principal y el foco de cada ciclo del estudio. Nuestros resultados sugieren que ninguna de las escalas de contexto que analizamos son completamente invariantes entre los países que participan en cada estudio, y por lo tanto las comparaciones entre países deben hacerse con precaución. Este artículo discute los niveles de equivalencia alcanzados por cada escala en cada estudio, así como el tipo de comparaciones que se pueden realizar dados estos resultados (e.g. comparación de los promedios nacionales de las escalas, comparación de relaciones o correlaciones entre las escalas evaluadas y otras variables, etc.).

Palabras clave: medición de la invarianza, medición de la equivalencia, TERCE, TIMSS, PISA, análisis factorial confirmatorio multigrupo, escalas de antecedentes socioeconómicos.

Abstract

Using data from international large-scale assessments (ILSA), we evaluate the issue of country-level model-data consistency of background socio-economic scales, as well as the invariance across countries. To that end, we use data from PISA, TERCE, and TIMSS, as they operationalize socio-economic status somewhat differently. As part of our analysis, we examine whether TERCE, a Latin American study – with measures that are regionally developed – exhibits better psychometric properties than measures that are designed to function across a larger and more diverse number of educational systems. We also examine TIMSS, a trends focused study – that has historically emphasized consistency and comparison. Finally, we include PISA which has the largest number of participants and has changed and conceptualized a great deal of its background questionnaire depending on the study's major domain and focus. Our findings suggest that none of the socio-economic background scales we analyzed are fully invariant in any of the three studies, and therefore comparisons across countries should be done with caution. The different levels of equivalence reached by each scale in each study and the type of comparisons that can be made given these results (e.g., comparison of average scale scores, comparison of relationships between the tested scales and other variables) are discussed in the full paper.

Key words: measurement invariance, measurement equivalence, TERCE, TIMSS, PISA, multi-group confirmatory factor analysis, socio-economic scales.

Introducción

Las evaluaciones internacionales de gran escala (ILSAs, por sus siglas en inglés) acerca del logro educativo, como el Programa para la Evaluación Internacional de Alumnos (PISA), el Estudio Internacional de Tendencias en Matemática y Ciencias (TIMSS) y el Tercer Estudio Regional Comparativo y Explicativo (TERCE), tienen múltiples propósitos. Desde monitorear sistemas educativos y realizar análisis comparativos, hasta proveer de información sobre lo que los estudiantes saben y pueden hacer. Los ILSAs ofrecen a las partes interesadas una oportunidad de entender el contexto y los factores asociados al aprendizaje, y entregan antecedentes acerca de los estudiantes, los profesores y las escuelas. No obstante, a medida que aumenta la participación en estas evaluaciones, se vuelve más difícil para las organizaciones adaptarlas y cumplir con las necesidades de un conjunto cada vez más heterogéneo de países. Por ejemplo, en la evaluación PISA 2015 participaron 62 países, 34 de ellos corresponden a los países miembros de la OCDE (que representan las mayores economías en el mundo), mientras que los 28 participantes restantes (denominados países asociados) correspondieron a un conjunto heterogéneo de economías y culturas, incluyendo sistemas educacionales como Túnez, Perú, Singapur y Shanghai, China. TIMSS enfrenta una situación similar. Finalmente, aunque las evaluaciones regionales como TERCE cuentan con una menor cantidad de participantes y –ostensiblemente– presentan menor heterogeneidad que las evaluaciones internacionales más globales, la diversidad de idiomas, economías y culturas persiste tanto entre los países participantes como dentro de ellos. Por ejemplo, el PIB per cápita de Chile triplica el de Bolivia y, aunque la mayoría de los países comparten el idioma español, en muchos de los países conviven distintos pueblos indígenas con diferentes lenguas nativas.

La mayoría de los ILSAs incluyen tanto una evaluación cognitiva como un conjunto de cuestionarios de contexto. Los cuestionarios se administran a estudiantes y, dependiendo de la evaluación, pueden incorporar a profesores, padres y directores de escuelas. En general, los cuestionarios de contexto tienen dos usos principales: (1) ayudar a contextualizar el sistema educacional evaluado; y (2) optimizar la estimación del logro académico en la población y un conjunto de subpoblaciones. Los beneficios de usar datos de contexto para mejorar la estimación del logro académico han sido ampliamente documentados

Recibido: 27-07-2017 Aceptado: 14-09-2018

(Mislevy, Beaton, Kaplan y Sheehan, 1992) y no constituyen el foco de este artículo. De hecho, algunos investigadores han destacado potenciales desafíos metodológicos asociados al amalgamiento de participantes, apuntando especialmente al modelo de estimación del logro y a si las comparaciones son razonables y válidas cuando los sistemas educativos difieren sustancialmente (Goldstein, 2004; Kreiner y Christensen, 2014; Mazzeo y von Davier, 2009; Oliveri y Ercikan, 2011). En parte como respuesta a estas y otras críticas, el proyecto PISA ha implementado ajustes, especialmente orientados hacia participantes de bajo rendimiento (e.g., incorporando ítems de menor dificultad a los instrumentos en países donde de espera bajo desempeño; OCDE, 2012). Estudios recientes han demostrado que estos tipos de ajustes son una manera promisoria de dar cuenta y enfrentar la heterogeneidad que está inevitablemente presente en la investigación comparativa (Rutkowski, Rutkowski y Zhou, 2016). Se ha realizado un gran esfuerzo para evaluar y asegurar la comparabilidad de las escalas de logro entre países (e.g. OECD, 2014; Schulz, Ainley y Frailon, 2011; UNESCO-OREALC, 2016) y a lo largo del tiempo (e.g. Gaviria y Covadonga, 2007). Por el contrario, se ha puesto mucho menor esfuerzo en diseñar escalas derivadas de los cuestionarios de contexto que consideren diferencias importantes entre los participantes (Rutkowski y Rutkowski, 2010).

Diversas investigaciones han demostrado empíricamente que el supuesto de equivalencia entre las escalas de contexto de los ILSAs a menudo no se sostiene, lo que compromete la comparabilidad (Caro, Sandoval-Hernandez y Lüdtke, 2016; Glas y Jehangir, 2014; Oliveri y von Davier, 2014). De esta manera, el objetivo de este artículo es doble: primero, mostrar un método para explorar tanto la consistencia de las escalas de contexto dentro de los países, como la equivalencia de estas escalas entre los países. El segundo objetivo es discutir los resultados de la aplicación de este método en el caso de las escalas de nivel socioeconómico de PISA, TIMSS y TERCE. Específicamente, exploramos los diferentes niveles de comparabilidad alcanzados por las escalas utilizadas en cada estudio para medir alguna forma de nivel socioeconómico (NSE) y discutimos el tipo de comparaciones que se pueden realizar en virtud de estos resultados (e.g., comparación de las medias de las puntuaciones, comparación de las relaciones entre las escalas evaluadas y otras variables).

Dado que no sería factible evaluar la comparabilidad de todas las escalas de contexto de los tres estudios, en este artículo nos enfocamos

Recibido: 27-07-2017 Aceptado: 14-09-2018

en las escalas desarrolladas por las organizaciones implementadoras para examinar alguna forma de NSE en tres estudios internacionales (PISA, TERCE y TIMSS). Decidimos usar estas escalas porque, en los estudios enfocados en identificar factores asociados a los resultados de aprendizaje (e.g., efectividad escolar y docente), NSE es la variable control que consistentemente muestra asociaciones más fuertes con el logro educativo. Además, existe un cuerpo importante de literatura focalizada especialmente en entender los mecanismos mediante los cuales los antecedentes socioeconómicos o el estatus socioeconómico de la familia están asociados al logro educativo (Buchmann, 2002).

Al examinar la equivalencia de estas escalas entre los países y al comparar los resultados de los diferentes estudios, podemos determinar si un diseño o enfoque de evaluación diferente resulta en distintos grados de comparabilidad. Estos tres estudios fueron elegidos intencionalmente ya que representan tres diseños de evaluación internacional diferentes. TERCE se escogió por representar un estudio regional, con preguntas que son desarrolladas a nivel regional y bajo el supuesto de que los desarrolladores del instrumento construyeron la escala para un grupo menor de participantes (Treviño, Fraser, Meyer, Morawietz, Inostroza y Naranjo, 2015). TIMSS se escogió por representar un estudio enfocado en tendencias – que ha históricamente puesto énfasis en la consistencia y la comparación de los cambios en las sociedades, constructos o participantes. Finalmente, incluimos PISA, que tiene el mayor número de participantes e históricamente ha demostrado interés en realizar cambios significativos a sus cuestionarios de contexto (OECD, 2016a).

Marco analítico

Nuestro marco analítico se sitúa a nivel general en la teoría de la medición y específicamente en la teoría y diseño de instrumentos de medición (e.g., van der Linden, 2005; Wilson, 2005). La teoría de los tests se focaliza en cómo un conjunto de respuestas observadas da cuenta de un constructo teórico, no observable. Dentro de los ILSAs, estas respuestas observables se obtienen a partir una prueba o instrumento (estandarizado), el que puede ser definido como "una técnica para relacionar algo que podemos observar en el mundo real (a veces denominado manifiesto u observable) con algo que medimos y que sólo existe como parte de una teoría (a

veces denominado latente o no observable)" (Wilson, p. 4). El diseño de instrumentos, o el proceso de desarrollar ítems que eliciten un constructo teórico no-observable, es un proceso iterativo. El supuesto que subyace a la teoría de los tests, y que también rige el diseño de los instrumentos, es que la relación entre los constructos teóricos y las respuestas observables a los ítems del instrumento es de tipo causal (Wilson, 2005). Esto es, el grado en que el participante presenta determinado constructo (o constructos en el caso multidimensional) provoca las respuestas a un conjunto de ítems. Debido a que no podemos observar el constructo directamente, el agente causal es latente y la medida (ítem) está construida para "inferir el constructo subyacente", permitiendo que el investigador solo pueda asumir causalidad (Wilson, p. 12).

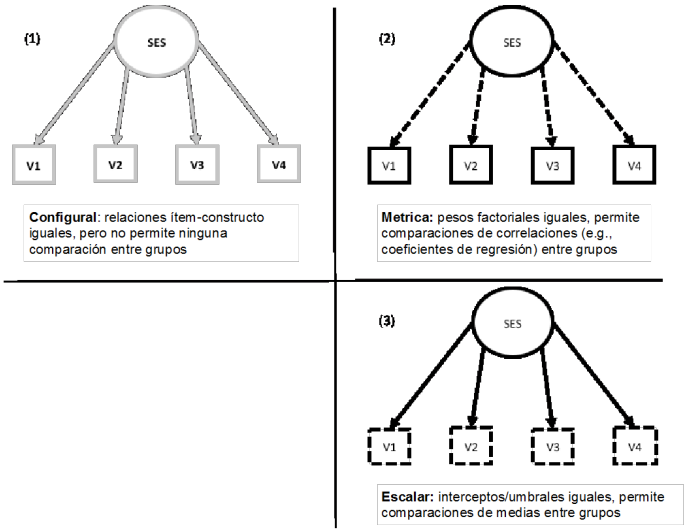
Para asumir una relación causal entre las respuestas observables y el rasgo latente, el desarrollo del instrumento requiere de un proceso riguroso de validación (Cronbach y Meehl, 1955; Messick, 1984). En este caso, un instrumento válido es aquel donde existe amplia evidencia que sugiera que los ítems están midiendo el constructo de interés para la población objetivo. La acumulación de esta evidencia es, en ocasiones, denominada proceso de validación (Shadish, Cook, y Campbell, 2002). Aunque el proceso de validación supone múltiples etapas, un aspecto importante es verificar la estructura correlacional del instrumento en la población de evaluados. Asumiendo que la estructura factorial se mantiene en esta población, el desarrollador del instrumento luego continúa con los siguientes pasos de la validación, como por ejemplo, estudios multirasgo-multimétodo.

Cuando un instrumento está diseñado para ser utilizado en múltiples poblaciones, como es el caso de los ILSAs, se requieren algunos pasos de validación adicionales para verificar que el instrumento opera de igual manera en todas las poblaciones. Brown (2015) menciona cuatro tipos de invarianza grupal para este propósito: a) formas iguales, b) pesos factoriales iguales, c) interceptos/umbrales iguales y d) varianzas residuales iguales (también conocidos como invarianza configural, métrica, escalar y estricta, respectivamente). El modelo de formas iguales es el tipo de invarianza más laxo y supone que la estructura de las relaciones ítem-constructo es idéntica en todos los grupos (ver representación gráfica en cuadrante 1 de la Figura 1). La estructura de pesos factoriales iguales se contrasta con la estructura previa (configural) y requiere que la varianza verdadera de cada ítem sea idéntica en todos los grupos (ver cuadrante

Recibido: 27-07-2017 Aceptado: 14-09-2018

2 de la Figura 1). Luego, el modelo de interceptos iguales, para ítems continuos, y umbrales iguales, en el caso de ítems discretos, supone que los ítems tienen la misma localización en el espacio latente (ver cuadrante 3 en la Figura 1). Finalmente, el modelo de varianzas residuales iguales, cuando se contrasta con el modelo de interceptos/umbrales iguales, indica que todos los ítems tienen la misma varianza en cada grupo, ya que el peso factorial y la varianza residual equivalen a la varianza total. Asegurar la invarianza del modelo de medición implica que el mismo constructo está siendo medido de igual manera en los diferentes grupos. Contar con evidencia de invarianza del modelo de medición no verifica automáticamente la relación causal entre el constructo y las respuestas; sin embargo, la incapacidad de demostrar equivalencia entre las poblaciones sugiere que el supuesto de causalidad no se mantiene. Es importante también mencionar que el nivel de invarianza requerido depende de los objetivos del análisis. Niveles diferentes de invarianza permiten realizar distintos tipos de comparaciones. La Figura 1 muestra un resumen de los tipos de comparación que se pueden realizar en distintos niveles de invarianza.

FIGURA 1. Diferentes niveles de invarianza y tipos de comparaciones permitidas en cada nivel



Nota: las líneas puntuadas representan la parte del modelo que se pone a prueba en cada nivel de invarianza.

Recibido: 27-07-2017 Aceptado: 14-09-2018

Método

Datos

Los datos de este estudio provienen de los último ciclos de tres ILSAs: TERCE, implementado por el Laboratorio Latinoamericano para la Evaluación de la Calidad de la Educación de la UNESCO (LLECE); TIMSS, conducido por la Asociación Internacional para la Evaluación del Logro Educativo (IEA) y PISA, realizado por la Organización para la Cooperación y Desarrollo Económicos (OCDE). Estos tres estudios, TERCE, TIMSS 2015 y PISA 2015 son los estudios comparativos internacionales más recientes que evalúan el logro de los estudiantes y recopilan información de un conjunto de actores educativos. Específicamente, PISA mide el logro de los estudiantes de 15 años en matemática, ciencias y lectura, con foco especial en la habilidad del estudiante para aplicar el conocimiento en contextos prácticos y situaciones de la 'vida cotidiana' (OECD, 2014, p. 24). En contraste al foco práctico de PISA, TIMSS y TERCE son instrumentos basados en el currículo y se enfocan en lo que los estudiantes han podido aprender en la escuela. TIMSS mide el logro en matemática de estudiantes de 4to y 8vo grado (Mullis, I.V.S., Martin, M.O., Foy, P., y Hooper, M., 2016), mientras que TERCE mide lectura, matemática y ciencias en 3er y 6to grado (Treviño, et al., 2015). En el presente estudio, utilizamos datos de los 72 sistemas educacionales que participaron en PISA 2015, los 44 sistemas educacionales que participaron en TIMSS 8vo grado, y los 16 sistemas educativos que participaron en TERCE 6to grado.

De cada estudio, seleccionamos intencionalmente una escala que cada organización implementadora construyó y liberó en sus bases de datos como un indicador de antecedentes familiares. Estas escalas provienen de los cuestionarios de estudiantes que se administran a cada participante después de haber completado la parte cognitiva de la evaluación. En PISA, el nivel socioeconómico se estima a través del Índice de Estatus Económico, Social y Cultural (ESCS, por sus siglas en inglés), obtenido a través de un conjunto de variables relacionadas con los antecedentes familiares del estudiante, por ejemplo, educación de los padres, ocupación de los padres, bienes que posee la familia en el hogar, posesiones culturales, y el número de libros disponibles en el hogar (OECD, 2016c, p.205). En TIMMS, utilizamos la Escala de Recursos Educativos en el Hogar (HERS, por sus siglas en inglés), que fue creada a partir de las

Recibido: 27-07-2017 Aceptado: 14-09-2018

respuestas de los estudiantes acerca de la disponibilidad de tres recursos: número de libros en el hogar, nivel educacional más alto alcanzado por los padres y un número de herramientas educativas en el hogar (Martin, Mullis, Hooper, Yin, Foy y Palazzo, 2016). En TERCE, utilizamos la Escala de Nivel Socioeconómico y Cultural de la Familia (ISECF), que proviene de los siguientes ítems: educación de los padres, ocupación de los padres, ingreso familiar y disponibilidad de diferentes posesiones y servicios en el hogar (UNESCO-OREALC, 2016). Debido a que los constructos teóricos de los tres estudios no son equivalentes, el objetivo de nuestro análisis fue examinar el grado en que las organizaciones implementadoras son capaces de crear una escala que sea comparable entre los países que participan en sus estudios, y no comparar la misma escala entre distintos estudios. La Tabla 1 muestra el conjunto de indicadores utilizados en cada estudio para medir los antecedentes socioeconómicos.

TABLA 1. Indicadores utilizados en cada estudio para construir una medida de antecedentes socioeconómicos.

Escala/Estudio	Ítem	Descripción
PISA: Índice de Estatus Económico, Social y Cultural (ESCS)[2]	1. Estatus ocupacional más alto de los padres (HISEI). 2. Nivel educacional más alto de los padres (PARED). 3. Posesiones en el hogar (HOMEPOS).	1. Los datos ocupacionales del padre y la madre se obtuvieron a partir de preguntas abiertas. Las respuestas se codificaron en códigos ISCO de cuatro dígitos y luego fueron agrupadas en un índice socioeconómico internacional de estatus ocupacional. 2. Nivel más alto de educación de cualquiera de los padres, recodificado en las siguientes categorías: (0) ninguno, (1) educación primaria, (2) segundaria baja, (3) secundaria alta vocacional / pre-vocacional, (4) secundaria alta general y/o post-secundaria no terciaria, (5) vocacional terciaria y (6) terciaria teóricamente orientada o postgrado. El índice corresponde al nivel ISCED más alto de cualquiera de los padres. 3. Los estudiantes reportaron la disponibilidad de 16 bienes en el hogar, incluyendo tres bienes específicos de cada país y el número de libros en la casa. Luego, se calculó un índice que resume todos bienes y posesiones del hogar utilizando un modelo TRI mediante WLE (logits) para las dimensiones latentes, las que fueron transformadas a escalas de media 0 y desviación estándar 1 (con pesos muestrales iguales).

(2) Estas variables se derivan consecuentemente de un conjunto de ítems individuales. Ver (OCDE, 2016c) para más detalles sobre el procedimiento llevado a cabo para construir esta escala.

Recibido: 27-07-2017 Aceptado: 14-09-2018

TIMSS: Recursos Educacionales en el Hogar (HERS)	1. Número de libros en el hogar. (BSBG04). 2. Número de apoyos para el estudio en el hogar. (BSDG06S). 3. Nivel más alto de educación de cualquier de los padres (BSDGED-UP).	1. Categorías de respuesta: (1) 0-10, (2) 11-25, (3) 26-100, (4) 101-200, (5) Más de 200. 2. Categorías de respuesta: (1) Ninguno, (2) Conexión a Internet o habitación propia, (3) Ambas. 3. Nivel educativo más alto de cualquiera de los padres, recodificado en las siguientes categorías: (1) primaria completa o algún grado de secundaria baja o no fue a la escuela, (2) secundaria baja completa, (3) secundaria alta completa, (4) educación post-secundaria completa, (5) universitaria o mayor completa.
TERCE: Escala de Nivel Socio-económico y Cultural de la Familia (ISECF)	1. Nivel educativo más alto de la madre (DQ-FIT09_02). 2. Nivel ocupacional más alto de la madre (DQ-FIT11_02). 3. Ingreso mensual del hogar (DQ-FIT12). 4. Material del piso de la casa (DQ-FIT14). 5. Servicio en el hogar (BIENES1). 6. Posesiones en el hogar (BIENES2). 7. Número de libros en la casa (DQ-FIT21).	1. Categorías de respuesta: (1) ninguno, (2) educación primaria, (3) mayor a educación primaria. 2. Categoría de respuesta: (1) Nunca ha trabajado fuera del hogar, (2) personal de limpieza, mantenimiento, construcción, agricultor, etc., (3) vendedor, operario de máquinas, conduce vehículos motorizados, etc., (4) trabajo administrativo, dueño de un negocio pequeño, (5) profesional, dueño de un negocio grande, a cargo de una división o área de una compañía, etc. 3. Ingreso declarado recodificado en deciles de ingreso del país con las siguientes categorías: (1) decil 1, (2) decil 2, (3) decil 3, (4) decil 4, (5) decil 5, (6) decil 6 a 10. 4. Categorías de respuestas: (1) tierra, (2) cemento o tablas de madera sin pulir., (3) Baldosas, cerámica o similar, (4) Parquet, madera pulida o piso alfombrado. 5. Los estudiantes reportaron la disponibilidad de 5 servicios en el hogar: desagüe o alcantarillado, recolección de basura, teléfono fijo, televisión por cable o satelital y conexión a internet. Luego, se utilizó análisis de componentes principales (ACP) para crear un índice resumen de los servicios en el hogar. 6. Los estudiantes reportaron el número de los siguientes bienes en el hogar: televisor, radio, computador, refrigerador, lavadora de ropa, celular, auto. Luego, se utilizó un análisis de componentes principales (ACP) para crear un índice que resume los ítems bienes en el hogar. 7. Categorías de respuesta: (1) niguno, (2) 10 o menos, (3) 11-20, (4) 21-30, (5) más de 31.

Fuente: OECD, 2016c; Martin, et al., 2016; UNESCO-OREALC, 2016.

Recibido: 27-07-2017 Aceptado: 14-09-2018

Estrategia analítica

Nuestra estrategia analítica consistió en dos etapas. Utilizamos primero un análisis factorial confirmatorio (AFC) con el objetivo de comprobar la estructura factorial del modelo utilizado para medir algún aspecto de antecedentes socioeconómicos[3] en cada estudio (i.e., TIMSS, PISA y TERCE). Se ajustó un modelo AFC para cada país en cada estudio. Luego, utilizamos un análisis factorial confirmatorio multigrupo (AFCMG) para verificar diferentes niveles de invarianza entre los países para cada escala en cada uno de los tres estudios descritos anteriormente. AFCMG (Jöreskog, 1971) es una de las técnicas más utilizadas para evaluar la invarianza de los modelos de medición (Billiet, 2003). AFCMG es una extensión del AFC que se utiliza para evaluar diferencias grupales en medias y covarianzas dentro de un modelo factorial común (Jöreskog, 1971); o como McGrath (2015) señala, para evaluar el ajuste global en múltiples grupos (sistemas educacionales en nuestro caso).

En la primera etapa, para evaluar el ajuste de cada modelo en cada país, utilizamos cuatro indicadores: la prueba de chi-cuadrado, el índice de ajuste comparativo (CFI), el índice de Tucker-Lewis (TLI) y la raíz del error cuadrático medio de aproximación (RMSEA). Adoptamos los puntos de cortes propuestos por Rutkowski y Svetina (2014) para realizar análisis en contextos donde la cantidad de grupos es grande y los tamaños muestrales son grandes y heterogéneos (e.g., muestras de ILSAs): ≤ .10 para el RMSEA, ≥ .95 para CFI y TLI. Aunque la prueba de chi-cuadrado no es considerada de utilidad en este contexto (Meade et al., 2008; Rutkowski y Svetina; 2014, Cheung y Rensvold, 2002), también reportamos su valor para analizar si las escalas de comportan de la manera esperada en todas las condiciones, donde los chi-cuadrado generalmente aumenta a medida que se especifican más restricciones en estos modelos.

Es importante comentar que en aquellos casos donde la escala de nivel socioeconómico está compuesta solamente de tres indicadores, como TIMSS y PISA, el modelo de un factor es un modelo saturado (i.e., no tiene grados de libertad). Como consecuencia, la evaluación

(3) Aunque la escala de TIMSS no es un índice de nivel socioeconómico, la Escala de Recursos Educacionales en el Hogar se utiliza comúnmente en las publicaciones de la IEA como una medida de aproximación a los antecedentes socioeconómicos de los estudiantes. Ver por ejemplo: Martin et al. 2013; Erberber et al., 2015; Trude y Gustafsson, 2016.

del ajuste no se puede realizar porque un modelo de tres indicadores tiene ajuste perfecto. De cualquier modo, de acuerdo con Brown (2015) estos "modelo[s] pueden evaluarse igualmente en términos de su interpretabilidad y del valor de sus parámetros estimados (e.g., magnitud de los pesos factoriales)" (pp. 71).

En la segunda etapa, con el objetivo de verificar la invarianza de las escalas de nivel socioeconómico entre los grupos (i.e., sistemas educativos), se ajustaron modelos AFCMG en todos los grupos simultáneamente dentro de cada estudio. Esto es, se ajustó un modelo AFCMG con todos los países participantes de TERCE, se ajustó un segundo modelo AFCMG para todos los países participantes de TIMSS, y un tercer modelo AFCMG se ajustó para todos los países participantes de PISA. De acuerdo a las prácticas comunes en este campo, ajustamos una serie de modelos anidados que van desde los modelos menos restrictivos a los más restrictivos. De esta manera, comenzamos evaluando cada escala desde el modelo configural, seguido por el modelo métrico y escalar. Aunque es posible evaluar el modelo de invarianza restrictiva o de varianzas residuales iguales (i.e., el cuarto nivel de invarianza) en la jerarquía propuesta por Brown (2015), la invarianza escalar es suficiente para realizar comparaciones significativas entre las medias latentes de los grupos (Marsh et al., 2010; Meredith 1993).

En esta segunda etapa, llevamos a cabo dos conjuntos de análisis. Primero, para examinar el comportamiento de los índices de ajuste de los AFCMG, los modelos se ajustaron para todos los países simultáneamente en cada estudio, donde la prueba de invarianza configural fue seguida por las pruebas de invarianza métrica y escalar. Siguiendo a Rutkowski y Svetina (2014), este primer conjunto de análisis fue denominado *medidas de ajuste global*. Evaluamos cada modelo (e.g., configural, métrico y escala) utilizando los mismos criterios presentados anteriormente. De esta manera, CFI y TLI debieron ser no menores a .95, y RMSEA no mayor a .10.

Luego, para verificar la factibilidad de los modelos de invarianza métrica y escalar, utilizamos ΔCFI, ΔTLI y ΔRMSEA entre modelos consecutivos más y menos restrictivos (configural vs métrico, y métrico vs escalar). Este segundo conjunto de análisis fue denominado *medidas de ajuste relativo*. Considerando los tamaños muestrales grandes y heterogéneos y el número relativamente grande de grupos (i.e., sistemas educacionales), seguimos la propuesta de Rutkowski y Svetina (2014). Para comprobar

Recibido: 27-07-2017 Aceptado: 14-09-2018

la invarianza métrica, estas diferencias debieron ser ΔCFI ≤ 0.020, ΔTLI ≤ 0.020 y ΔRMSEA 0.030. Para comprobar la invarianza escalar, las diferencias debieron ser ΔCFI ≤ 0.010, ΔTLI ≤ 0.010 y ΔRMSEA 0.010.

Resultados

Primero, para cada estudio/escala y para cada país, mostramos los resultados generales de los modelos AFC que describen el grado en que los indicadores empíricos se ajustan a los constructos teóricos propuestos por cada estudio. En segundo lugar, para cada estudio/escala mostramos los resultados del análisis multigrupo y las pruebas de invarianza entre todos los países participantes de cada estudio.

Etapa 1: Análisis por país. Comenzamos nuestro análisis con AFCs para cada país en cada estudio. Como las escalas de TIMSS y PISA tienen solo tres ítems, solo existe un único conjunto de parámetros que se ajustan y reproducen los datos (Harrington, 2009). Por esta razón, en vez de presentar una tabla con los índices de ajuste (que incluiría sólo valores constantes), seguimos el procedimiento propuesto por Miranda y Castillo (2018) y presentamos un gráfico que ilustra los pesos factoriales estandarizados de cada ítem. Esto nos permite evaluar los modelos en relación a la magnitud de los pesos factoriales de cada ítem (Brown, 2015). La Figura 1 presenta los pesos factoriales de la escala de PISA, la Figura 2 contiene los de la escala de TIMSS y la Figura 3 los de la escala del TERCE. Para mantener la consistencia, presentamos el gráfico de pesos factoriales estandarizados del modelo del TERCE, aun cuando este está identificado (gl > 0).

En las Figuras 1, 2 y 3, cada punto representa el peso factorial estandarizado de cada ítem en un país determinado, mientras que la línea horizontal que cruza cada punto representa el intervalo de confianza al 95%. Incluimos una línea vertical en el eje que representa un peso factorial de 0.5, ya que este valor se considera el mínimo aceptable para un peso estandarizo en AFC (Hair et al., 2006).

FIGURA 2. Pesos factoriales estandarizadas para cata ítem que compone el SES en PISA, TIMSS y TERCE.

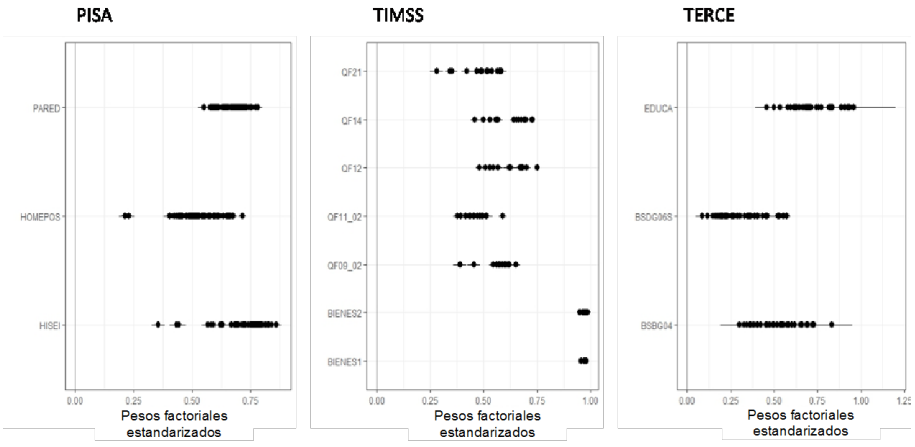


TABLA 2. Número de países con pesos factoriales bajo/sobre 0.5 por indicador/estudio.

Estudio	Indicador	Peso factorial	
		< 0.5	**> 0.5**
PISA	HISEI	3	65
	PARED	0	68
	HOMPOS	25	43
TIMSS	BSBG04	16	28
	BSDG06S	37	7
	EDUCA	1	43
TERCE	QF09_02	0	16
	QF21	7	9
	QF11_02	12	4
	QF12	1	15
	QF14	1	15
	BIENES1	0	16
	BIENES2	0	16

Recibido: 27-07-2017 Aceptado: 14-09-2018

La Figura 2 presenta los pesos factoriales de cada indicador utilizado para medir antecedentes socioeconómicos en los tres estudios, mientras que la Tabla 3 muestra el número de países que presentan pesos factoriales menores y mayores a 0.5 en cada indicador, para cada estudio considerado. Como se observa, en PISA los indicadores PARED y HISEI muestran pesos factoriales mayores a 0.5 en la mayoría de los países (68 y 65, respectivamente. Ver Tabla 2). El indicador HOMEPOS presenta pesos factoriales bajo 0.5 en 25 países, e incluso valores menores a 0.25 en dos países (ver Tabla 2). En TIMSS, como se observa en la Figura 2, el indicador EDUCA es el único que tiene pesos factoriales mayores a 0.5 en la mayoría de los países (43, ver Tabla 2). Los otros dos indicadores presentan mayor variación entre países. Por ejemplo, el indicador BSBG04 presenta pesos factoriales menores a 0.5 en 16 países, y el indicador BSDG06S presenta pesos factoriales bajo 0.5 en 37 países, con valores bajo 0.25 en casi la mitad de estos (ver Tabla 2). Finalmente, en TERCE, la Figura 2 muestra que ninguno de los indicadores presenta pesos factoriales menores a 0.25. Los ítems QF21 y QF11_02, sin embargo, presenta pesos factoriales menores a 0.5 en algunos países (7 y 12, respectivamente. Ver Tabla 2). Particularmente BIENES1 y BIENES2, presentan pesos factoriales mayores a 0.5 en los 16 países participantes del estudio (ver Tabla 2).

Hasta ahora, hemos ilustrado que las medidas de nivel socioeconómico analizadas y su configuración en los países presentan variaciones importantes entre los estudios. Nuestros resultados sugieren que, entre las escalas de antecedentes socioeconómicos analizadas, la escala de TERCE es la que tiene menor variación en su configuración entre los países y la que tiene mejor ajuste, seguida por las escalas de PISA y TIMSS.

Etapa 2: análisis multigrupo. Las pruebas de invarianza indicaron que se alcanzaron diferentes niveles de invarianza en los tres estudios analizados. En el caso de PISA, a partir de la información entregada por los pesos factoriales, es heurísticamente razonable pensar que la estructura de la escala es similar en todos los países (ver Figura 2). Los tres indicadores tuvieron parámetros relativamente estables entre los países, con pesos factoriales por sobre 0.50 en la mayoría de los indicadores y la mayoría de los países. Sólo el índice HOMEPOS mostró algunos pesos factoriales menores a 0.25 (Qatar y Emiratos Árabes Unidos). Como se observa en la Tabla 3, el modelo métrico mostró índices de

ajuste por sobre los criterios de corte, mientras que el modelo escalar mostró índices de ajuste bajo los criterios establecidos (ver Tabla 3). Sin embargo, las medidas de *ajuste relativo* indicaron que no se alcanzó el nivel de invarianza métrica ni el escalar.

TABLE 3. Medidas de ajuste global de los AFCMG para cada nivel de invarianza

	PISA				TIMSS				TERCE			
Modelo	X^2	*CFI*	*TLI*	*RMSEA*	X^2	*CFI*	*TLI*	*RMSEA*	X^2	*CFI*	*TLI*	*RMSEA*
Configural	0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000	4124.600	0.979	0.968	0.070
Métrico	5951.480	0.974	0.961	0.077	3195.150	0.863	0.790	0.077	6216.210	0.968	0.965	0.072
Escalar	61932.700	0.728	0.793	0.177	19723.990	0.146	0.672	0.096	16862.100	0.910	0.925	0.107

En TIMSS, la información de ajuste global (e.g., pesos factoriales) muestra que el modelo base de invarianza configural posee una gran dispersión de los pesos factoriales (ver Figura 2). Concretamente, solo en siete países encontramos pesos factoriales sobre 0.05. y cerca de un tercio de los países (e.g., Canadá, Hungría, Irlanda, Italia, Japón y Kuwait) tuvieron pesos factoriales bajo 0.25. Los otros dos indicadores son relativamente estables entre los países. El modelo métrico mostró índices de *ajuste global* por debajo de los criterios de corte y, en consecuencia, no obtuvimos evidencia para afirmar que se alcanzó el nivel de invarianza métrica o escalar (ver Tabla 3). Asimismo, los índices de *ajuste relativo* sugieren que no se alcanzó ni el nivel de invarianza métrica ni el escalar (ver Tabla 4).

Finalmente, el TERCE mostró buenos índices de *ajuste global* en el caso de los modelos configural y métrico, pero estuvieron fuera del rango aceptable en el caso del modelo escalar (ver Tabla 3 y Figura 2 para pesos factoriales del modelo configural). En relación a las medidas de ajuste relativo, las comparaciones entre los modelos configural y métrico entregan evidencia de invarianza métrica (ver Tabla 4).

En resumen, nuestros análisis mostraron índices de ajuste que no entregan evidencia de invarianza métrica o escalar en el caso de las escalas de antecedentes económicos utilizadas en TIMSS y PISA, mientras que la escala del TERCE sí cuenta con evidencia de invarianza métrica y escalar.

Recibido: 27-07-2017 Aceptado: 14-09-2018

TABLA 4. Medidas de ajuste relativo de los AFCMG para cada nivel de invarianza.

	PISA			TIMSS			TERCE		
Model	X^2*diff.*	Δ*CFI*	Δ*RMSEA*	X^2*diff.*	Δ*CFI*	Δ*RMSEA*	X^2*diff.*	Δ*CFI*	Δ*RMSEA*
Métrico	5951.476	0.026	0.077	3126.102	-0.137	0.077	2086.991	-0.011	0.002
Escalar	44692.380	0.246	0.100	17830.425	-0.717	0.019	7523.642	-0.058	0.035

Discusión

A nivel elemental, los cuestionarios de contexto están compuestos por múltiples instrumentos, algunos de ellos diseñados para medir constructos hipotéticos (e.g., estatus socioeconómico). Como estos constructos hipotéticos no se pueden medir directamente, son las respuestas a un determinado conjunto de ítems de los cuestionarios de contexto las que sirven como indicadores indirectos del constructo, el que se operacionaliza a través de un modelo de medición. Como los constructos son teóricos, fenómenos no-observables, el acto de verificar o validar el instrumento es un proceso extremadamente importante que cae dentro del marco de la teoría moderna de la medición (e.g., van der Linden, 2005; Wilson 2005). Las ideas centrales para desarrollar y validar dichos instrumentos son: comenzar con un constructo bien definido, diseñar un conjunto de ítems que se supone elicitan dicho constructo y luego verificar que el modelo de medición propuesto sea consistente con los datos generados por esos ítems. Cuando un modelo propuesto no se ajusta a las respuestas de uno o más ítems, este o estos se deben revisar o reemplazar. En otras palabras, el desarrollo de constructos no debiese ser un ejercicio posterior, en el que se exploran ítems para definir constructos posibles, los que luego se mapean nuevamente en alguna teoría. Al menos, no debiese ser un ejercicio lineal que termina con el mejor intento (aunque insuficiente) para ajustar los datos empíricos a una teoría determinada.

Además, solo luego de que existe evidencia suficiente que sugiera que el instrumento ajusta bien en una población, es que puede comenzar la tarea de evaluar la factibilidad del instrumento para realizar comparaciones entre poblaciones. Un procedimiento común para verificar la invarianza de los modelos de medición entre países es evaluar la equivalencia de la estructura de covarianzas, medias y varianzas residuales. Al examinar el

grado de igualdad del modelo de medición, somos capaces de verificar estadísticamente el supuesto de comparabilidad de las escalas. Si el supuesto se mantiene entonces habría evidencia estadística de que las escalas pueden ser comparadas razonablemente. Pero, como ocurrió en el presente análisis, los constructos no siempre son comparables, lo que sugiere que los atributos medidos no son iguales en los países.

Existe una serie de explicaciones plausibles cuando se descubre que las escalas no son invariantes entre poblaciones o culturas. Lo primero, y lo más importante, es que es absolutamente posible que el constructo teórico no sea pertinente o que haya sido pertinente alguna vez, pero debido a cambios en la sociedad, ocurre que ya no tiene relevancia. Por supuesto, si el constructo no es relevante la escala no debería reportarse y el constructo antiguo debiese reemplazarse por un constructo nuevo. En situaciones donde existe gran apoyo para un constructo universal, existen otras posibles razones por la que una escala puede no ser invariante, incluyendo:

1) El constructo se puede medir, pero el marco teórico no es pertinente;
2) El constructo se puede medir, el marco teórico es pertinente, pero los indicadores se operacionalizaron incorrectamente.
3) El constructo se puede medir, el marco teórico es pertinente, pero no existen indicadores universales.

En relación al primer punto, una estrategia viable y relativamente sencilla de realizar consiste en revisar el marco teórico utilizado para operacionalizar el constructo en cuestión. El segundo punto tiene un poco más de matices. Operacionalizar constructos incorrectamente podría ocurrir por una serie de motivos. Por ejemplo, el marco de referencia que constituye el fundamento para medir el NSE de los niños debiese incluir el ingreso del hogar, una pregunta que es difícil obtener de manera confiable a partir de niños jóvenes. De esta manera, deben recolectarse otros indicadores más indirectos del ingreso del hogar. Aunque obtener indicadores de posesiones en el hogar en estas circunstancias sea lo más razonable, estas variables pueden no reflejar precisamente el ingreso del hogar. Por ejemplo, es posible que, con el auge de la lectura online, el número de libros en el hogar ya no represente ni el capital económico ni el NSE. En el tercer escenario, el constructo existe, pero los indicadores necesarios para medir el constructo difieren entre países o regiones. Nuevamente, el NSE es un ejemplo útil para ilustrar este punto.

Recibido: 27-07-2017 Aceptado: 14-09-2018

La mayoría de la teoría académica define el NSE como un constructo universal. Para fundamentar este constructo, aquellos investigadores interesados en medir el NSE en estudios internacionales podrían aplicar un marco conceptual universal. Pero a pesar de la aceptación de la teoría del NSE, los indicadores que representan el constructo pueden diferir por país. Por ejemplo, un indicador confiable de capital económico familiar en EEUU podría ser si el niño tiene una habitación propia o si la familia ha tenido vacaciones en el extranjero. En cambio, un indicador relativamente deficiente sería si la familia tiene una cortadora de césped. Por el contrario, una cortadora de césped es un signo importante de capital económico en Hong Kong o Singapur, dada la relativa falta de tierras donde hacer crecer el césped. Hasta ahora, la pregunta sobre la universalidad de los indicadores pero no de la teoría se mantiene – existe algún conjunto de indicadores que permita diferenciar internacionalmente y de manera confiable aquellos niños de altos y bajos recursos? Claro es que, dado el comportamiento de las mediciones que se examinaron aquí, así como también otras investigaciones similares (Caro, Sandoval-Hernandez y Lüdtke, 2016; Rutkowski y Rutkowski, 2017), aún queda mucho trabajo por hacer.

Un camino posible es relajar el requerimiento de que los constructos deban ser definidos de manera idéntica en todos los sistemas evaluados. Aunque PISA, por ejemplo, ha permitido la inclusión de ítems de capital económico específicos de cada país, estos no son incluidos en el modelo unidimensional del NSE. En cambio, son tratados como variables observables y específicas del país que luego se transforman en combinaciones lineales de variables para construir una medida de estatus sociocultural. Sin embargo, tales combinaciones lineales no corresponden a variables latentes (Bentler, 1982) y no tienen una estructura hipotetizada. Estas aproximaciones no miden algo, sino que corresponden a meros ejercicios de reducción de datos.

Es posible, en cambio, ajustar modelos de variables latentes que adhieran al supuesto de invarianza parcial, donde se permite incluir ítems únicos y estimar sus parámetros. Investigaciones previas han mostrado que, aunque operacionalizar estos constructos requiere más esfuerzo, esta parece ser una manera promisoria de mejorar la consistencia de los datos de los modelos entre países, manteniendo la comparabilidad.

Rutkowski y Rutkowski (2017) concentraron sus esfuerzos en la región Nórdica, que es relativamente homogénea, y demostraron que se

requiere investigar más para desarrollar medidas únicas de cada país que funcionen adecuadamente. Aunque sin duda esto requerirá un esfuerzo significativo por parte de los países participantes, también permitirá que estos incorporen matices locales y culturales, propios de sus contextos, a escalas posibles de comparar en estudios internacionales.

TERCE entrega otra posible solución. Dado que es una evaluación regional que se centra en países de lenguajes, culturas y economías similares (en comparación a PISA y TIMSS), TERCE debería ser capaz, con mayor dedicación, de diseñar y administrar cuestionarios que se ajusten mejor a poblaciones específicas. En el presente artículo, nuestros resultados indican que el TERCE fue capaz de desarrollar una escala de antecedentes socioeconómicos que es comparable a nivel métrico, lo que supera a su contraparte en TIMSS. A pesar de todo, ningún estudio mostró invarianza escalar, donde las medias latentes pudiesen ser válidamente comparadas entre países. En otras palabras, en TERCE, TIMMS y PISA, existe evidencia estadística para sugerir que el indicador de antecedentes socioeconómicos no es comparable entre distintas culturas. Peor aún, tanto en PISA como en TIMSS, las escalas no cumplen con los estándares básicos de calidad en muchos de los sistemas educacionales participantes. Dado esto, los análisis que utilizan los promedios de las escalas de nivel socioeconómico en cualquiera de estos estudios producen resultados, al menos, cuestionables.

Estos resultados tienen implicancias directas sobre las políticas educativas y la investigación. Por ejemplo, los estudios que estiman la proporción de estudiantes resilientes[4] en un grupo de países y luego realizan comparaciones entre países son un clásico ejemplo de esta práctica (e.g. OECD, 2011; Erberber, Stephens, Mamedova, Ferguson y Kroeger, 2015). Además, lo mismo puede decirse de cualquier estudio internacional comparativo que use el índice de estatus económico, social y cultural (ESCS) de PISA o la Escala de recursos educativos en el hogar (HERS) de TIMSS como variable de control en un modelo de regresión[5]. Nuestros hallazgos suponen una amenaza importare a la validez de estas escalas y cualquier análisis futuro debería alertar a los lectores de estas amenazas.

---

(4) Comúnmente definidos como estudiantes con bajo NSE y algo nivel de logro académico.

(5) De acuerdo a nuestros resultados, este tipo de comparación serían válidas usando la Escala de Nivel Socioeconómico y Cultural de la Familia (NSCF), ya que esta alcanzó el nivel de invarianza métrica.

Recibido: 27-07-2017 Aceptado: 14-09-2018

Conclusión

Las escalas de los cuestionarios de contexto tienen un rol importante para ayudar a explicar el logro educativo. De hecho, algunas escalas han cobrado vida propia y muchas veces operan fuera de los resultados de logro. Por ejemplo, escalas como *bullying*, compromiso de los estudiantes y compromiso cívico son importantes para las políticas educativas y también interesantes cuando no se relacionan con el logro. Aunque utilizamos el NSE, una escala común a las tres evaluaciones, como ejemplo para este estudio, se puede realizar un análisis similar en cualquier estudio que pretenda utilizar escalas de alguna ILSA para realizar estudios transculturales. Además, como se demostró en este artículo, se quiere más trabajo para mejorar las medidas internacionales. Al menos, las escalas de contexto de los ILSAs requieren un proceso de validación tan riguroso como el de las escalas de logro (OCDE, 2014). Tal proceso contribuiría a prevenir el reporte de escalas que no son comparables entre los países participantes.

Como sugirieron nuestros resultados, al adoptar un foco regional más riguroso en el desarrollo de los cuestionarios, los ILSAs podría mejorar la comparabilidad de determinados constructos, tales como el NSE. Concretamente, dos posibles caminos son: mejorar el desarrollo regional de los cuestionarios y entregar mayor financiamiento para el desarrollo de ILSAs regionales y sus cuestionarios. Podría argumentarse que el Estudio Internacional de Educación Cívica y Formación Ciudadana (ICCS), incluyendo sus módulos regionales, representa el primer camino y el TERCE representa el segundo de estos posibles modelos. En cada caso, sin embargo, un marco conceptual claro, que se exprese en escalas regionales específicas, sigue ausente y necesitaría ser desarrollado en extenso. En el caso de ILSAs de mayor escala como PISA y TIMSS, recomendamos, al menos, diversificar la composición cultural de las partes/instituciones a cargo de los marcos de referencia internacionales. Por ejemplo, el grupo experto que supervisó al comité del marco de referencia y los instrumentos de PISA 2015 incluyó once miembros provenientes principalmente de economías desarrolladas de la OCDE (más de la mitad de EEUU y Alemania). La composición del comité claramente no representaba la composición cultural extremadamente diversa de los participantes de PISA (OCDE, 2016b). Finalmente, en los casos donde los países y grupos de países encuentran que el marco

de referencia está mal especificado, los miembros deberían trabajar con las organizaciones desarrolladoras para realizar ajustes a los marcos de referencia y las escalas. Si eso no es posible, entonces los participantes deberían solicitar a las organizaciones de ILSAs la exclusión del país en cualquier reporte de esa escala. Por supuesto, este compromiso trae consigo un costo alto; sin embargo, el costo de publicar y utilizar malas escalas podría ser aún mayor.

Referencias bibliográficas

Bentler, P. M. (1982). Confirmatory factor analysis via non-iterative estimation. A fast inexpensive method. *Journal of Marketing Research*, *25A*(5), 309-318.

Billiet, J. (2003). Cross-cultural equivalence with structural equation modeling. En J. Harkness, F. Van de Vijver and P. Mohler (Eds.), *Cross-cultural survey methods (*pp. 247-264). NJ: John Wiley and Sons.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. New York: Guildford Press.

Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. En A. C. Porter and A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 150–197). Washington, DC: National Academy Press.

Caro, D. H., Sandoval-Hernández, A. and Lüdtke, O. (2016). Cultural, social, and economic capital constructs in international assessments: an evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*(3), 433-450.

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. DOI: 10.1207/S15328007SEM0902_5

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Erberber, E., Stephens, M., Mamedova, S., Ferguson, S., and Kroeger, T. (2015). Socioeconomically disadvantaged students who are academically successful: Examining academic resilience cross-

nationally. *IEA's Policy Brief Series*, No. 5, Amsterdam: IEA. Recuperado de http://www.iea.nl/policy_briefs.html

Gaviria Soto, J. L., Biencinto López, M. C., and Navarro Asencio, E (2007). Invarianza de la estructura de covarianzas de las medidas de rendimiento académico en estudios longitudinales en la transición de Educación Primaria a Secundaria. *Revista de Educación*, 348, 153-173.

Glas, C., and Jehangir, K. (2014). Modeling country-specific differential item functioning. En L. Rutkowski, M. von Davier, and D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: Chapman and Hall / CRC Press.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy and Practice*, *11*(3), 319–330. DOI: 10.1080/0969594042000304618

Harrington, D. (2009). *Confirmatory Factor Analysis*. Oxford: Oxford University Press.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 408-426.

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

Martin, M.O., Mullis, I.V.S., Hooper, M., Yin, L., Foy, P. and Palazzo, L. (2016). Creating and Interpreting the TIMSS 2015 Context Questionnaire Scales. En M. O. Martin, I. V. S. Mullis, and M. Hooper (Eds.) *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center / IEA.

Marsh, H.W., Ludtke O., Muthen, B., Asparouhov, T, Morin, A.J.S., et al. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychol. Assess. 22*(3), 471–91.

McGrath, R. E. (2015). Measurement Invariance in Translations of the VIA Inventory of Strengths. *European Journal of Psychological Assessment*. On-line advanced publication. DOI: 10.1027/1015-5759/a000248

Mazzeo, J., and von Davier, M. (2009). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB.* Paris: OECD.

Messick, S. (1984). The Psychology of Educational Measurement. *Journal of Educational Measurement*, *21*(3), 215–237.

Miranda, D. and Castillo, J. C. (2018). Measurement model and invariance testing of scales measuring egalitarian values in ICCS 2009. En A. Sandoval-Hernandez, M. M. Isac and D. Miranda (Eds.) *Teaching Tolerance in a Globalized World*. Cham: Springer International Publishing

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika 58*(4), 525–43

Mullis, I.V.S., Martin, M.O., Foy, P., and Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center / IEA.

OECD. (2011). *Against the Odds: Disadvantaged Students who Succeed in School*. Paris: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing.

OECD. (2016a). *PISA 2015 Assessment and Analytical Framework.* Paris: OECD Publishing.

OECD. (2016b). PISA 2015 Background questionnaires. Annex A (pp. 129–196). En *PISA 2015 Assessment and Analytical Framework*. Paris: OECD Publishing.

Oliveri, M. E., and Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education*, *24*(4), 349–366. DOI: 10.1080/08957347.2011.607063

Oliveri, M. E., and von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333.

Oliveri, M. E., and von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*(1), 1–21. DOI: 10.1080/15305058.2013.825265

Rutkowski, L. and Rutkowski, D. (2010). Getting it "better": The importance of improving background questionnaires in International

Recibido: 27-07-2017 Aceptado: 14-09-2018

Large-Scale Assessment. *Journal of Curriculum Studies, 42*(3), 411–430. DOI: 10.1080/00220272.2010.487546

Rutkowski, L. and Rutkowski, D. (2017). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 0*(0), 1–14. DOI: 10.1080/00313831.2016.1261044

Rutkowski, L., Rutkowski, D. and Zhou, Y. (2016). Parameter estimation methods and the stability of achievement estimates and system rankings: Another look at the PISA model. *International Journal of Testing, 16*(1), 1–20.

Rutkowski, L., y Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31-57.

Schulz, W., Ainley, J. and Fraillon, J. (Eds.). (2011). *ICCS 2009 Technical Report*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Treviño, E., Fraser, P., Meyer, A., Morawietz, L., Inostroza, P. and Naranjo, E. (2015). *Informe de Resultados TERCE. Factores Asociados*. Santiago: Oficina Regional de Educación para América Latina y el Caribe (OREAL/UNESCO).

UNESCO-OREALC. (2016). *Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo, TERCE*. Santiago: Oficina Regional de Educación para América Latina y el Caribe (OREAL/UNESCO).

Van Der Linden, W. J. (2005). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement*, *42*(3), 283-302.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Dirección de contacto: Andrés Sandoval-Hernández. University of Bath, Faculty of Humanities & Social Sciences, Department of Education. University of Bath, Claverton Down, Bath, BA1 6TP, United Kingdom. **E-mail:** A.Sandoval@bath.ac.uk

Back to the drawing board: Can we compare socioeconomic background scales?[1]

Pensémoslo de nuevo: ¿Podemos comparar las escalas de antecedentes socioeconómicos?

DOI: 10.4438/1988-592X-RE-2019-383-400

Andrés Sandoval-Hernandez
University of Bath
David Rutkowski
Indiana University
University of Oslo
Tyler Matta
Pearson
University of Oslo
Daniel Miranda
Pontificia Universidad Católica de Chile, MIDE UC
Centro de Estudios de Conflicto y Cohesión Social

Abstract

Using data from international large-scale assessments (ILSA), we evaluate the issue of country-level model-data consistency of background socio-economic scales, as well as the invariance across countries. To that end, we use data from PISA, TERCE, and TIMSS, as they operationalize socio-economic status somewhat differently. As part of our analysis, we examine whether TERCE, a Latin American study – with measures that are regionally developed – exhibits better psychometric properties than measures that are designed to function across a larger and more diverse number of educational systems. We also examine TIMSS, a trends focused

(1) This paper was supported by the Norwegian Research Council Grant #255246 and by the Center of Social Conflict and Cohesion Studies —COES CONICYT/FONDAP N°°15130009

study – that has historically emphasized consistency and comparison. Finally, we include PISA which has the largest number of participants and has changed and conceptualized a great deal of its background questionnaire depending on the study's major domain and focus. Our findings suggest that none of the socio-economic background scales we analyzed are fully invariant in any of the three studies, and therefore comparisons across countries should be done with caution. The different levels of equivalence reached by each scale in each study and the type of comparisons that can be made given these results (e.g., comparison of average scale scores, comparison of relationships between the tested scales and other variables) are discussed in the full paper.

Key words: measurement invariance, measurement equivalence, TERCE, TIMSS, PISA, multi-group confirmatory factor analysis, socio-economic scales.

Resumen

Utilizando datos de evaluaciones internacionales de gran escala, evaluamos la consistencia y la invarianza de las escalas de antecedentes socioeconómicos de los estudiantes entre los países participantes en estos estudios. Para ello, utilizamos las medidas de antecedentes socioeconómicos desarrolladas por PISA, TERCE y TIMSS, ya que cada estudio operacionaliza esta medida de manera diferente. Como parte de nuestro análisis, examinamos si la escala de TERCE, un estudio latinoamericano – con medidas que fueron desarrolladas con enfoque regional – exhibe mejores propiedades psicométricas que otras medidas que fueron diseñadas para funcionar en un número mayor y más diverso de sistemas educativos. Adicionalmente, examinamos la escala de TIMSS, un estudio enfocado en tendencias – que históricamente ha enfatizado la comparabilidad y consistencia. Finalmente, incluimos también la escala de PISA, que cuenta con el mayor número de participantes y que, en cierta medida, ha cambiado y conceptualizado sus cuestionarios de contexto dependiendo el dominio principal y el foco de cada ciclo del estudio. Nuestros resultados sugieren que ninguna de las escalas de contexto que analizamos son completamente invariantes entre los países que participan en cada estudio, y por lo tanto las comparaciones entre países deben hacerse con precaución. Este artículo discute los niveles de equivalencia alcanzados por cada escala en cada estudio, así como el tipo de comparaciones que se pueden realizar dados estos resultados (e.g. comparación de los promedios nacionales de las escalas, comparación de relaciones o correlaciones entre las escalas evaluadas y otras variables, etc.).

Palabras clave: measurement invariance, measurement equivalence, TERCE, TIMSS, PISA, multi-group confirmatory factor analysis, socio-economic scales.

Received: 27-07-2017 Accepted: 14-09-2018

Introduction

International large-scale assessments (ILSAs) of educational achievement such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Third Regional Comparative and Explanatory Study (TERCE) serve multifold purposes. From system monitoring and benchmarking to providing national participants and researchers with information about what students know and can do. ILSAs offer stakeholders an opportunity for understanding the context and correlates of learning in a number of areas as well as provide important background information on students, teachers, and schools. As national participation in these assessments grows, however, it is becoming more difficult for testing organizations to tailor assessments to meet the needs of a diverse set of participants. For example, in the 2015 PISA cycle, all 34 OECD member countries (representing the largest economies in the world, excepting China) participated in PISA, with the remaining 38 participants (termed *partner* systems) comprised of a heterogeneous mix of economies and cultures, including educational systems such as Tunisia, Peru, Singapore, and Shanghai, China. A similar situation is also faced in TIMSS. Finally, although regional assessments such as TERCE have fewer participants and – ostensibly – less heterogeneity than more global international studies, language, economic and cultural diversity among and within TERCE participants persists. For example, Chile's GDP per capita is over three times that of Bolivia and although the majority of countries share Spanish as a common language many participating countries include indigenous populations with a variety of mother tongues.

Most ILSAs include both a cognitive assessment and a set of background questionnaires. The questionnaires are administered to students and, depending on the assessment, can measure others such as teachers, parents and school leaders. In general, the background questionnaires have two primary uses: (1) to help contextualize the assessed educational system; and (2) to optimize population and sub-population achievement estimation. The benefits of using background data to help estimate achievement are well documented (Mislevy, Beaton, Kaplan, and Sheehan, 1992) and are not the focus of this paper. In fact, potential methodological challenges associated with an amalgamation of participants have been highlighted by a number of researchers, who have pointed especially

to the achievement estimation model and whether comparisons are sensible and valid when systems differ dramatically (Goldstein, 2004; Kreiner and Christensen, 2014; Mazzeo and von Davier, 2009; Oliveri and Ercikan, 2011). Partly in response to these and other criticisms, the PISA project has implemented accommodations, especially targeted toward lower performing participants (e.g., incorporating easier items into the test for countries with low expected performance; OECD, 2012). Recent research has demonstrated that these types of accommodations are a promising way of acknowledging and dealing with the heterogeneity that is necessarily present in cross-cultural research (Rutkowski, Rutkowski and Zhou, 2016). Significant work has been done to ensure comparability of achievement scales across countries (e.g. OECD, 2014; Schulz, Ainley and Frailon, 2011; UNESCO-OREALC, 2016) and across time (e.g. Gaviria and Covadonga, 2007). In contrast, much less effort is spent on designing scales derived from the background questionnaires that can account for vast differences among participants (Rutkowski and Rutkowski, 2010).

Empirically, research has shown that the assumption of equivalent background scales in ILSAs is often violated, leading to compromised comparability (Caro, Sandoval-Hernandez and Lüdtke, 2016; Glas and Jehangir, 2014; Oliveri and von Davier, 2014). As such, the objective of this paper is twofold: First, to demonstrate a method to explore both within-county data consistency and equivalence across countries on background scales. Second, to discuss the results of the application of this method to the socio-economic scales of PISA, TIMSS and TERCE. More specifically, we explore the different levels of equivalence reached by the scales used in each study to measure some form of socio-economic status (SES) and discuss the type of comparisons that can be made given these results (e.g., comparison of average scale scores, comparison of relationships between the tested scales and other variables).

It would not be feasible to evaluate the equivalence of all the background scales of the three studies, for this reason in this paper we focus on the scales developed by the testing organizations to examine some form of SES in three international studies (PISA, TERCE, and TIMSS). We decided to use these scales because, in the studies focused on identifying factors associated with learning outcomes (e.g. school and teacher effectiveness), SES is the control variable that consistently shows a stronger association with educational achievement. Furthermore, there is an important body of literature specifically focused on understanding

Received: 27-07-2017 Accepted: 14-09-2018

the mechanisms by which socio-economic background or family socio-economic status is associated with academic achievement (Buchmann, 2002).

By examining the equivalence of these scales between countries and comparing the findings across studies, we can determine if different assessment designs or approaches result in different degrees of comparability. The three studies were purposely chosen because they represent three different designs of international assessment. TERCE was chosen to represent a regional study – with measures that are regionally developed and with the assumption that test developers were able to focus the scale for a smaller group participants (Treviño, Fraser, Meyer, Morawietz, Inostroza and Naranjo, 2015). TIMSS was chosen to represent a trend focused study – one that has historically emphasized consistency and comparison over changes in societies, constructs, or participants. Finally, we include PISA, which has the largest number of participants and has historically been willing to make significant changes to its background questionnaires (OECD, 2016a).

Analytical framework

Our analytical framework is broadly situated within measurement theory and more specifically within ideas of test theory and design (e.g., van der Linden, 2005; Wilson, 2005). Test theory is focused on how a set of observed responses can map onto a theoretical, unobservable construct. Within ILSAs these observed responses are elicited from a (standardized) test or instrument, which may be defined as "a technique of relating something we observed in the real world (sometimes called manifest or observed) to something we are measuring that only exists as part of a theory (sometimes called latent or unobserved)" (Wilson, p. 4). Instrument design, or the process of developing items that elicit an unobserved theoretical construct, is an iterative process. An underlying assumption in test theory, which governs instrument design, is that the relationship between the theoretical construct and the observed responses to the items that make up the instrument is a causal one (Wilson, 2005). That is, a respondent's level on a particular construct (or constructs in the multidimensional case) causes their responses for a set of items. Because we cannot observe the construct directly, the causal agent is latent and the

Received: 27-07-2017 Accepted: 14-09-2018

measure is left to "infer the underlying construct" allowing the researcher to only assume causality (Wilson, p. 12).

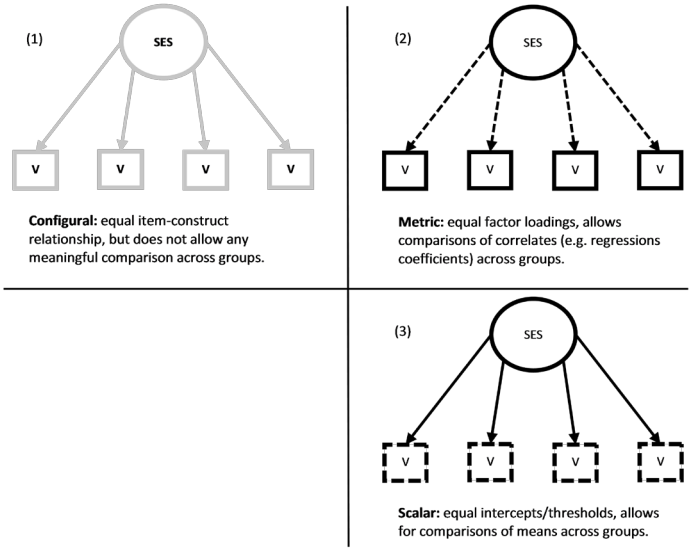
In order to assume a causal relationship between the observed responses and the latent trait, the instrument development requires a rigorous validation process (Cronbach and Meehl, 1955; Messick, 1984)1955; Messick, 1984. In this case, a valid instrument is one where there is ample evidence that suggests the items are measuring the intended theoretical construct for the selected population. The accumulation of the evidence is sometimes referred to as the validation process (Shadish, Cook, and Campbell, 2002). Although the validation process includes multiple steps, one important aspect is to test the correlational structure of the instrument within the intended population of respondents. Assuming the factor structure holds for the given population, the instrument designer then takes additional steps toward validation, such as multitrait-multimethod studies.

When an instrument is intended to be used with multiple populations, as is intended with ILSAs, further validation is required to ensure the instrument operates in the same way across all populations. Brown (2015) outlined four types of group invariance for this purpose: a) equal form, b) equal loadings, c) equal intercepts/ thresholds, and d) equal residual variances (also known as configural, metric, scalar and strict invariance, respectively). The equal form is the most lenient type of invariance and means that the structure of the item-construct relationship is identical across all groups (see quadrant 1 in Figure 1 for a graphical representation). The test of equal loadings builds upon the previous structure and requires that the true score variance in each item is identical across all groups (see quadrant 2 in Figure 1). Next, equal intercepts for continuous items and equal thresholds for discrete items demonstrates that the items have the same locations in the latent space (see quadrant 3 in Figure 1). Finally, equal residual variances, when built upon equal intercepts/thresholds, indicates that all items have the same amount of variance across each group since the loading plus the residual variance equals the total variance. Ensuring measurement invariance indicates that the same construct is being measured in the same way across different groups. Evidence of measurement equivalence does not automatically validate the causal relationship between the construct and respondent; however, an inability to demonstrate equivalence across populations suggests that the assumption of causality between the respondent and

Received: 27-07-2017 Accepted: 14-09-2018

construct does not hold. It is also important to mention that the level of invariance required depends on the objectives of the analysis. Different levels of invariance allow for different types of comparisons. See Figure 1 for a summary of the types of comparisons allowed by the different levels of invariance.

FIGURE 1. Different levels of invariance and types of comparisons allowed in each level



Note: The dotted lines represent the part of the model that is being tested in each invariance level

Methodology

Data

The data for this study has been sourced from the latest cycles of three major ILSAs: TERCE, managed by the UNESCO's Latin American Laboratory for Assessment of the Quality of Education (LLECE); TIMSS, managed by the International Association for the Evaluation of Educational Achievement (IEA), and PISA managed by the Organisation for Economic Cooperation and Development (OECD). All three studies,

TERCE, TIMSS 2015 and PISA 2015 are the most recent international comparative studies that assess student achievement and gather information from a range of educational stakeholders. Specifically, PISA measures student achievement in mathematics, science and reading of 15 years' old students, with a focus on students' ability to apply knowledge in practical contexts and 'everyday life' situations (OECD, 2014, p. 24). In contrast to PISA's focus on practical situations, TIMSS and TERCE are curriculum-based tests and focus on what students had an opportunity to learn in school. TIMSS measures students' achievement in mathematics and science at 4th and 8th grade (Mullis, I.V.S., Martin, M.O., Foy, P., and Hooper, M., 2016), while TERCE measures reading, mathematics and science at 3rd and 6th grades (Treviño, et al., 2015). In this study, we used data from the 72 education systems participating in PISA 2015, from the 44 education systems participating TIMSS 8th grade, and from the 16 education systems that participated in TERCE 6th grade.

For each of the studies, we purposely selected a scale that each testing organisation has constructed and included in their released database as a proxy measure of family background. These scales result from the student questionnaire that is administered to each participant after they complete the cognitive portion of the assessment. In PISA, student's socio-economic status is estimated by the index of economic, social and cultural status (ESCS), which is derived from several variables related to students' family background: parents' education, parents' occupation, a number of home possessions that can be taken as proxies for material wealth and cultural possessions, and the number of books available in the home (OECD, 2016c, p.205). In TIMSS, we used the Home Educational Resources Scale (HERS), which was created based on students' responses concerning the availability of three resources: number of books in the home, highest level of parental education, and number of home study supports (Martin, Mullis, Hooper, Yin, Foy and Palazzo, 2016). In TERCE we used the Family Socioeconomic and Cultural Status Scale (FSCS), which was derived from the following items: parental education, parental occupation, family income, and availability of different home possessions and services (UNESCO-OREALC, 2016). Although the theoretical constructs are not the same across studies, our analyses are intended to examine the extent to which testing organizations are able to create scales that are comparable among the countries that participate in their studies, rather than compare

Received: 27-07-2017 Accepted: 14-09-2018

the same scale across studies. Table 1 shows the set of indicators used in each study to measure socioeconomic background.

TABLE 1. Indicators used in each study to construct a proxy measure of socioeconomic background

Scale / Study	Item	Description
PISA: Index of economic, social and cultural status (ESCS)[2]	1. Highest occupational status of parents (HISEI). 2. Highest educational level of parents (PARED). 3. Home possessions (HOMEPOS).	1. Occupational data for both the student's father and mother were obtained from responses to open-ended questions. The responses were coded to four-digit ISCO codes and then mapped to the international socio-economic index of occupational status. 2. Highest level of education od either parent, recoded onto the following categories: (0) none, (1) primary education, (2) lower secondary, (3), vocational/pre-vocational upper secondary, (4) general upper secondary and/or non-tertiary post-secondary, (5) vocational tertiary and (6) and/or theoretically oriented tertiary and post-graduate. The index corresponds to the higher ISCED level of either parent. 3. Students reported the availability of 16 household items at home, including three country-specific household items and the amount of books at home. Then a summary index of all household and possession items was calculated using IRT modelling with WLEs (logits) for the latent dimensions which were transformed to scales with an average of 0 and a standard deviation of 1 (with equally weighted samples).
TIMSS: Home Educational Resources (HERS)	1. Number of books in the home. (BSBG04). 2. Number of home study supports (BSDG06S). 3. Highest level of education of either parent (BSDGEDUP).	1. Response categories: (1) 0-10, (2) 11-25, (3) 26-100, (4) 101-200, (5) More than 200. 2. Response categories: (1) None, (2) Internet connection or own room, (3) Both. 3. Highest level of education of either parent, recoded onto the following categories: (1) finished primary or some lower-secondary or did not go to school, (2) finished lower-secondary, (3) finished upper-secondary, (4) finished post-secondary education, (5) finished university of higher.

(2) These variables are in turn derived from a set of individual items. See (OECD, 2016c) for more details on the procedure followed to construct this scale.

Received: 27-07-2017 Accepted: 14-09-2018

TERCE: Family Socioeconomic and Cultural Status Scale (FSCS)	1. Highest level of education of the mother (DQFIT09_02). 2. Highest occupational level of the mother (DQFIT11_02). 3. Monthly household income (DQFIT12). 4. Material the floor is made of in the home (DQFIT14). 5. Services in the home (BIENES1). 6. Home possessions (BIENES2). 7. Number of books in the home (DQFIT21).	1. Response categories: (1) none, (2) primary education, (3) more than primary education 2. Response categories: (1) has never worked out of the household, (2) cleaning, maintenance, construction, farmer, etc. (3) sales, operated machines, driver, etc., (4) administrative, owner of smal business, (5) professional, owner of medium/large business, managerial, etc. 3. Income declared recoded into country-income deciles with the following categories: (1) decil 1, (2) decil 2, (3) decil 3, (4) decil 4, (5) decil 5, (6) decil 6 to 10. 4. Response categories: (1) dirt, (2) cement or non-polished wood, (3) tiles or similar, (4) carpet, parquet or polished wood. 5. Students reported the availability of 5 services in the home: drainage, garbage collection, telephone landline, cable TV and internet connection . Then a summary index of all service items was calculated using principal component analysis (PCA). 6. Students reported the amount of the following household items in the home: TV, .radio, PC, refrigerator, washing machine, smart phone, car Then a summary index of all home possessions was calculated using principal component analysis (PCA). 7. Response categories: (1) none, (2) 10 or less, (3) 11-20, (4) 21-30, (5) more than 31

Source: OECD, 2016c; Martin, et al., 2016; UNESCO-OREALC, 2016.

Analytical strategy

Our analytical strategy consisted of two main steps. We first used confirmatory factor analysis (CFA) to test the factor structure of the model used in each study (i.e. TIMSS, PISA and TERCE) to measure some aspect socioeconomic background[3]. One CFA model was fit separately for each country in each study. Then, we used multi-group confirmatory factor analysis (MGCFA) to test for different levels of invariance across countries for each scale in each of the three studies described above. MGCFA (Jöreskog, 1971) is one of the most commonly used techniques to

(3) Even though the TIMSS scale is not a socioeconomic index, the Home Educational Resources Scale is commonly used in IEA's publications as measure to proxy student socioeconomic background. See for example: Martin et al., 2013; Erberber, et al., 2015; Trude and Gustafsson, 2016.

Received: 27-07-2017 Accepted: 14-09-2018

assess measurement invariance (Billiet, 2003). MGCFA is a straightforward extension of CFA that is used to evaluate group differences in means and covariances within a common factor model (Jöreskog, 1971); or as McGrath (2015) puts it, to evaluate overall model fit across multiple groups (education systems in our case).

In the first step, in order to evaluate the goodness of fit for each model in each country, we used four measures: the chi-squared test, comparative fit index (CFI), Tucker-Lewis index (TLI) and root mean square error of approximation (RMSEA). We followed the cut-off points proposed by Rutkowski and Svetina (2014) for the analyses in contexts where the number of groups is large and the sample sizes are large and varied (e.g. ILSA samples): ≤ .10 for RMSEA, ≥ .95 for CFI and TLI. Although the chi-square test is not considered to be useful in this context (Meade et al., 2008; Rutkowski and Svetina; 2014, Cheung and Rensvold, 2002), we also report chi-square statistics in order to analyse if the scales behave as expected across all conditions in that these values are generally larger as more constraints were placed on these models.

It is important to note that for those cases in which the socioeconomic scale is formed with only three indicators, such as in TIMSS and PISA, the one-factor solution is just-identified (i.e. does not have degrees of freedom). As a consequence, the evaluation of fit indexes is not possible because a three-indicator model has a perfect fit. In any case, according to Brown (2015) these "model[s] can still be evaluated in terms of the interpretability and strength of its parameter estimates (e.g., magnitude of factor loadings)" (pp. 71).

In the second step, in order to test for the invariance of the socioeconomic scales across groups (i.e. education systems), MGCFA models were fitted to all groups simultaneously within each study. That is, one MGCFA model was fit to all countries participating in TERCE, a second MGCFA model was fit to all countries participating in TIMSS, and a third MGCFA model was fit to all countries participating in PISA. According to the common practice in the field, we conducted a series of nested tests that proceed from least to most restrictive models. In this way, we started by testing each scale for configural invariance, followed by metric and scalar invariance. Although it is possible to test for strict invariance or equal residual variances (i.e. the fourth level of invariance) in the hierarchy proposed by Brown (2015), scalar invariance is sufficient

for meaningful comparison of latent means across groups (Marsh et al., 2010; Meredith 1993).

We carried out two sets of analyses within this second step. First, in order to examine the performance of MGCFA fit measures, MGCFA models were fit to all countries simultaneously, by study, where the test of configural invariance was followed by the tests of metric and scalar invariance. Following Rutkowski and Svetina (2014), we term this first set of analyses as *overall fit measures.* We evaluate each model (i.e. configural, metric and scalar) using the same criteria presented above. That is, CFI and TLI should be no smaller than .95, and RMSEA should be no larger than .10.

Then, in order to test the plausibility of metric and scalar invariance, we use ΔCFI, ΔTLI, and ΔRMSEA between more and less restrictive models (configural vs. metric, and metric vs. scalar). We term this second set of analyses as *relative fit measures.* Considering the large and varying sample sizes and the relative high number of groups (i.e. educational systems), we use the approach proposed by Rutkowski and Svetina (2014). For the test to show metric invariance, these differences must be ΔCFI ≤ 0.020, ΔTLI ≤ 0.020, and ΔRMSEA 0.030. For the test to show scalar invariance, these differences must be ΔCFI ≤ 0.010, ΔTLI ≤ 0.010, and ΔRMSEA 0.010.

Results

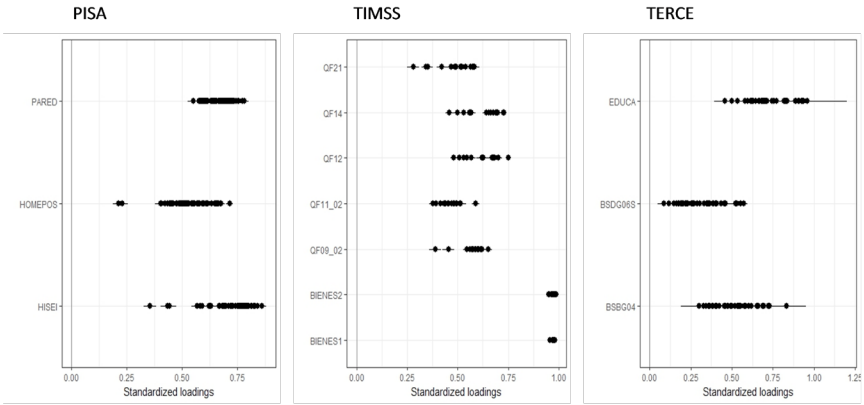
First, we show the general results regarding the extent to which the empirical indicators correspond to the theoretical constructs proposed by each study, tested by CFA procedure for each study/scale and for each country. Second, we show the results of the multi-group analyses and the test for measurement invariance for each study/scale across the education systems participating in each study.

Step 1: Single-country analysis. We start our analysis with separate CFAs for each country in each study. Because the scales for TIMSS and PISA have only three items, there is one unique set of parameters that perfectly fit and reproduce the data (Harrington, 2009). For this reason, instead of presenting a table with the fit indexes (which would include only constant values), we follow the follow the approach proposed by Miranda and Castillo (2018) and present a graph showing the standardised

Received: 27-07-2017 Accepted: 14-09-2018

factor loadings for each item. This allows us to evaluate the models in terms of the magnitude of the factor loadings of each item (Brown, 2015). Figure 1 shows the factor loadings for the PISA scale, Figure 2 for the TIMSS scale, and Figure 3 for the TERCE scale. Even when the model for TERCE is over-identified (df > 0), for consistency purposes, we present the graph with the standardised factor loadings. In Figures 1, 2 and 3, each dot represents the standardised factor loading of each item in one given country, and the horizontal line crossing each dot represents the confidence interval at the 95% level. We marked a vertical line at a 0.5 factor loading as this can be considered the minimum acceptable value for standardised loading in CFA (Hair et al., 2006).

FIGURE 2. Standardised factor loadings for each item composing SES in PISA, TIMSS and TERCE



Received: 27-07-2017 Accepted: 14-09-2018

TABLE 2. Number of countries under/above 0.5 factor loading by indicator/study

Study	Indicator	Factor loading	
		< 0.5	> 0.5
PISA	HISEI	3	65
	PARED	0	68
	HOMPOS	25	43
TIMSS	BSBG04	16	28
	BSDG06S	37	7
	EDUCA	1	43
TERCE	QF09_02	0	16
	QF21	7	9
	QF11_02	12	4
	QF12	1	15
	QF14	1	15
	BIENES1	0	16
	BIENES2	0	16

Figure 2 presents the factor loadings of each indicator used for measuring socioeconomic background in the three studies, and Table 3 shows the number of countries in which the factor loadings of each indicator are below and above 0.5 for each of the studies considered. As can be observed, for PISA, the indicators PARED and HISEI show factor loadings above 0.5 values in most countries (68 and 65, respectively. See Table 2). The indicator HOMEPOS presents 25 countries with factor loadings below 0.5, and even two countries with values under 0.25 (see Table 2). For TIMSS, as can be observed in Figure 2, the indicator EDUCA is the only one that has factor loadings above 0.5 for most countries (43, see Table 2). The other two indicators present higher variations across countries. For instance, the indicator BSBG04 presents 16 countries with factor loadings under 0.5, and the indicator BSDG06S presents factor loadings under 0.5 in 37 countries, with about half of these countries with values under 0.25 (see Table 2). Finally, in TERCE, Figure 2 shows that none of the indicators presents factor loadings under 0.25. The items QF21 and QF11_02, however, present some countries with factor loadings under 0.5 (7 and 12, respectively. See Table 2). Particularly BIENES1

Received: 27-07-2017 Accepted: 14-09-2018

and BIENES2, present factor loadings above 0.5 in all the 16 countries participating in the study (see Table 2).

So far, we have illustrated that the analysed socioeconomic background measures and their configuration across countries show important variations among studies. Our results suggest that among the socioeconomic background scales analysed, the TERCE scales is the one with least variations in its configuration across countries and the one with the best fit, followed by the PISA and TIMSS scales.

Step 2: Multi-group analysis. The test of measurement invariance indicated different levels of invariance for the three analysed studies. In the case of PISA, using the information from the factor loadings, heuristically, it is reasonable to assume that the structure of the scale is similar across countries (see Figure 2). The three indicators had relatively stable estimates across countries, with factor loadings over 0.50 for most indicators in most countries. Only the HOMEPOS index showed some factor loadings under 0.25 (in Qatar and the United Emirates). As can be observed in Table 3, the metric model showed fit indices over the cut-off criteria, while the scalar model showed fit indices under the established criteria (see Table 3). However, the *relative fit* measures indicate that neither metric nor scalar invariance was achieved (see Table 4).

TABLE 3. MGCFA overall fit measures for each level of invariance

	PISA				**TIMSS**				**TERCE**			
Model	X^2	*CFI*	*TLI*	*RMSEA*	X^2	*CFI*	*TLI*	*RMSEA*	X^2	*CFI*	*TLI*	*RMSEA*
Configural	0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000	4124.600	0.979	0.968	0.070
Metric	5951.480	0.974	0.961	0.077	3195.150	0.863	0.790	0.077	6216.210	0.968	0.965	0.072
Scalar	61932.700	0.728	0.793	0.177	19723.990	0.146	0.672	0.096	16862.100	0.910	0.925	0.107

For TIMSS, the *overall fit* information (e.g. factor loadings) shows that baseline model of configural invariance indicates high dispersion in factor loadings (see Figure 2). Particularly, in only seven countries we found factor loadings for the 0.05, and about one-third of the countries (e.g. Canada, Hungary, Ireland, Italy, Japan, Kuwait) had factor loadings under 0.25. The other two indicators are relatively stable across countries.

The metric model showed *overall* fit indices under the cut-off criteria, and as a consequence, there is no evidence to suggest that metric or scalar invariance was achieved (see Table 3). Similarly, the results of the *relative* fit indices suggest that neither metric nor scalar invariance was achieved (see Table 4).

Finally, the TERCE study showed good *overall* fit indices for the configural and metric models, but fit indices were out of the acceptable range for the scalar model (see Table 3 and Figure 2 for the factor loadings of the configural model). Regarding the *relative* fit measures, the comparison between the configural and metric models provide evidence of metric invariance (see Table 4).

In summary, our analyses resulted in fit indices that did not provide evidence of metric or scalar invariance for the socioeconomic background scales used in TIMSS and PISA, while the TERCE scale showed evidence of metric and scalar invariance.

TABLE 4. MGCFA relative fit measures for each level of invariance

	PISA			**TIMSS**			**TERCE**		
Model	X^2*diff.*	Δ*CFI*	Δ*RMSEA*	X^2*diff.*	Δ*CFI*	Δ*RMSEA*	X^2*diff.*	Δ*CFI*	Δ*RMSEA*
Metric	5951.476	0.026	0.077	3126.102	-0.137	0.077	2086.991	-0.011	0.002
Scalar	44692.380	0.246	0.100	17830.425	-0.717	0.019	7523.642	-0.058	0.035

Discussion

At a basic level, background questionnaires are made of up of multiple instruments, parts of which are intended to measure a hypothetical construct (e.g., socioeconomic status). Because hypothetical constructs cannot be measured directly, answers to select background questionnaire items serve as an indirect indicator of the construct, operationalized through a measurement model. Because constructs are theoretical, unobservable phenomena, the act of verifying or validating the instrument is an extremely important process, a process that falls under modern test theory (e.g., van der Linden, 2005; Wilson 2005). The core idea of

Received: 27-07-2017 Accepted: 14-09-2018

developing and validating such instruments is to begin with a well-defined construct, design a set of items that are assumed to elicit that construct, and then test if the proposed measurement model is consistent with the data generated from those items. When a proposed model does not fit data from one or more items, the items are revised or replaced. In other words, construct development should generally not be a posthoc exercise, where item responses are explored for viable constructs, which are then mapped back onto some theory. At least, it should not be a linear exercise that is finished with the best (but insufficient) attempt to fit the empirical data to a given theory.

Furthermore, only after sufficient evidence suggests the instrument is fitting well within a population can the task of evaluating the suitability of the instrument for cross-population comparisons begin. A common approach to testing measurement invariance across countries is to test the equality of the measurement model's covariance structure, means, and residual variances across countries. By examining the measurement model's degree of equality, we are able to statistically test the assumption of scale comparability. If the assumption holds then there would be statistical evidence that the scales can be compared sensibly. But, as was the case in the current examination, constructs are not always comparable suggesting that the traits being measured are not the same across cultures.

When scales are found to be non-invariant across populations or cultures a number of plausible explanations exist. First, and foremost, it is completely possible that a theorized construct is simply wrong or that it was once correct but changes in society have occurred so that the specified construct is no longer relevant. Of course, if the construct is not relevant the scale should not be reported and the old construct should be abandoned for a new construct. In situations where there exists strong theoretical backing for a universal construct there are other possible reasons that a scale may be found to be non-invariant to include:

1) The construct can be measured but the framework is incorrect;
2) The construct can be measured, the framework is correct, but the indicators are being operationalized incorrectly;
3) The construct can be measured, the framework is correct, but there are no universal indicators.

Received: 27-07-2017 Accepted: 14-09-2018

In terms of the first point, a viable and relatively straightforward treatment is that the framework used to operationalize the construct in question needs to be revised. The second point is a bit more nuanced. Operationalizing constructs incorrectly could happen for a variety of reasons. For example, the framework that stands as the foundation for measuring children's SES should include household income, which is a difficult question to reliably obtain from young children. Thus, other, more indirect indicators of household income must be collected. Potential alternatives could include the number of televisions, bedrooms, or books in the home. Although these might be the most reasonable variables to collect under the circumstance, measures of home possessions may not accurately reflect household income. For example, it is possible that with the dawn of e-readers the number of books in the home no longer represents either wealth or SES. In the third scenario, the construct exists but the indicators needed to measure that construct differ between countries or regions. Again, SES is a useful example to illustrate this point. The majority of academic theory may define SES as a universal construct. In support of the construct, a universal framework might be applied by researchers wishing to measure SES internationally. But regardless of the accepted theory of SES, the indicators that represent the construct may differ by country. For example, a reliable indicator of family wealth in the U.S. might be whether the child has a room of their own or whether they have taken an international vacation. In contrast, a relatively poor indicator would be if the family has a lawnmower. In contrast, a lawnmower is a strong signal of wealth in Hong Kong or Singapore, given the relative lack of land on which to grow grass. The question of universal indicators but not a universal theory remains –is there a set of indicators that can reliably differentiate between well– and poorly-resourced children internationally? Clearly, given the performance of the measures examined here as well as similar research (Caro, Sandoval-Hernandez and Lüdtke, 2016; Rutkowski and Rutkowski, 2017) much work remains to be done.

One possible way forward is to relax the requirement that constructs should be identically defined across measured systems. Although PISA, as one example, has allowed for the inclusion of country-specific wealth items, these are not subsumed under a single latent variable model of SES. Rather, they are treated as observed, country-specific variables that are formed into linear combinations of variables to make up a measure

Received: 27-07-2017 Accepted: 14-09-2018

of socio-cultural status. However, such linear combinations are not latent variables (Bentler, 1982), and have no hypothesized structure. Such approaches are not measuring anything but are merely exercises in data reduction.

It is possible to instead fit latent variable models that adhere to an assumption of partial invariance, whereby unique items and unique item parameters are allowed. Previous research has shown that, although more work needs to be done to operationalize this construct, it is a promising way to improve model-data consistency across countries while maintaining comparability. Importantly, Rutkowski and Rutkowski (2017) concentrated their efforts on the relatively homogeneous Nordic region and showed that more research needs to be done to develop well-functioning country-specific measures. Although this certainly will necessitate meaningful work on the part of participating countries, it will allow for participants to incorporate the local cultural nuances of their local context into internationally comparable scales.

TERCE provides another possible solution. As a regional assessment that focuses on similar language groups, cultures, and economies (when compared to PISA and TIMSS), with more focus TERCE should be able to design and administer questionnaires that are better tailored to a specific population. In the current manuscript, our results indicate that TERCE was able to develop a socioeconomic background scale that is comparable at the metric level which is better than its TIMSS counterpart. Regardless no study had acceptable scalar invariance where latent means can be validly compared across countries. In other words, for TERCE TIMSS and PISA, there is statistical evidence to suggest that the socio-economic background indicator is not cross-culturally comparable. Even worse, in both PISA and TIMSS the scales are not meeting basic quality standards within many participating educational systems. As such, analyses that use mean values of the socioeconomic scales on any of these studies will produce findings that are questionable, at best. These findings have direct policy and research implications. For example, studies that estimate the share of resilient students[4] in a group of countries and then make cross-national comparisons are a classic example of such practice (e.g. OECD, 2011; Erberber, Stephens, Mamedova, Ferguson and Kroeger, 2015). Furthermore, the same can be said about any international comparative

(4) Commonly defined as students with low SES and high academic achievement.

study that uses the PISA index of economic, social and cultural status (ESCS) or the TIMSS' home educational resources scale (HERS) as a control variable in a regression model[5]. Our findings pose a serious threat to the validity of these scales and any future analysis should caution readers to these threats.

Conclusion

Scales from the background questionnaire play an important role in helping to explain educational achievement. In fact, certain scales have taken a life of their own and often operate outside of the achievement results. For example, scales such as bullying, student engagement, and civic engagement are important to policy and interesting in absence of their relationship to achievement. Although, we used SES, a scale common to all three assessments, as an example for this study a similar analysis should be completed for any study that wishes to use scales from ILSAs for cross cultural comparison. Further, as demonstrated in this paper, substantial work needs to be done to improve measures internationally. At the very least, ILSA background scales require a validation process as rigorous as the achievement scales (OECD, 2014). Such a process would go a long way in preventing reporting scales that are not comparable across participating countries.

As hinted at by our results, by embracing a more rigours regional focus to questionnaire development, ILSAs might improve the comparability of certain constructs such as SES. More specifically, improving regional development of questionnaires or further funding the development of regional ILSAs and there questionnaires are two possible ways forward. It could be argued that the IEA's International Civic and Citizenship Study (ICCS), with its regional modules, represents the former and TERCE represents the latter of these possible models. In each case, however, a clear framework that maps directly onto regional specific scales is currently lacking and would need to be fully developed. In the case of larger ILSAs such as PISA and TIMMS, we recommend, at the very

(5) According to our results this type of comparison would be valid when using the TERCE's family socioeconomic and cultural status scale (FSCS) as this scale reached metric invariance (see tables 4 and 5).

Received: 27-07-2017 Accepted: 14-09-2018

least, diversifying the cultural makeup of those stakeholders who oversee the current international frameworks. For example, the expert group that oversaw the PISA 2015 questionnaire framework and instruments committee included eleven members that were mostly from highly developed OECD economies (over half from the U.S. and Germany). The committee composition clearly did not represent the extremely diverse cultural makeup of PISA participants (OECD, 2016b). Finally, in cases where countries or groups of countries find that the framework is miss-specified, members should work with the testing organizations to make adjustments to the framework and scales. If that is not possible, then participants should ask ILSA organizations not to include their country in any scale reporting. This engagement, of course, comes with a cost; however, publishing and using poor scales can be even more costly.

References

Bentler, P. M. (1982). Confirmatory factor analysis via non-iterative estimation. A fast inexpensive method. *Journal of Marketing Research*, *25A*(5), 309-318.

Billiet, J. (2003). Cross-cultural equivalence with structural equation modeling. En J. Harkness, F. Van de Vijver and P. Mohler (Eds.), *Cross-cultural survey methods (*pp. 247-264). NJ: John Wiley and Sons.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. New York: Guildford Press.

Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. En A. C. Porter and A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 150–197). Washington, DC: National Academy Press.

Caro, D. H., Sandoval-Hernández, A. and Lüdtke, O. (2016). Cultural, social, and economic capital constructs in international assessments: an evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*(3), 433-450.

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. DOI: 10.1207/S15328007SEM0902_5

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Erberber, E., Stephens, M., Mamedova, S., Ferguson, S., and Kroeger, T. (2015). Socioeconomically disadvantaged students who are academically successful: Examining academic resilience cross-nationally. *IEA's Policy Brief Series*, No. 5, Amsterdam: IEA. Recuperado de http://www.iea.nl/policy_briefs.html

Gaviria Soto, J. L., Biencinto López, M. C., and Navarro Asencio, E (2007). Invarianza de la estructura de covarianzas de las medidas de rendimiento académico en estudios longitudinales en la transición de Educación Primaria a Secundaria. *Revista de Educación*, 348, 153-173.

Glas, C., and Jehangir, K. (2014). Modeling country-specific differential item functioning. En L. Rutkowski, M. von Davier, and D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: Chapman and Hall / CRC Press.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy and Practice*, *11*(3), 319–330. DOI: 10.1080/0969594042000304618

Harrington, D. (2009). *Confirmatory Factor Analysis*. Oxford: Oxford University Press.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 408-426.

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

Martin, M.O., Mullis, I.V.S., Hooper, M., Yin, L., Foy, P. and Palazzo, L. (2016). Creating and Interpreting the TIMSS 2015 Context Questionnaire Scales. En M. O. Martin, I. V. S. Mullis, and M. Hooper (Eds.) *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center / IEA.

Received: 27-07-2017 Accepted: 14-09-2018

Marsh, H.W., Ludtke O., Muthen, B., Asparouhov, T, Morin, A.J.S., et al. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychol. Assess. 22*(3), 471–91.

McGrath, R. E. (2015). Measurement Invariance in Translations of the VIA Inventory of Strengths. *European Journal of Psychological Assessment*. On-line advanced publication. DOI: 10.1027/1015-5759/a000248

Mazzeo, J., and von Davier, M. (2009). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB*. Paris: OECD.

Messick, S. (1984). The Psychology of Educational Measurement. *Journal of Educational Measurement*, *21*(3), 215–237.

Miranda, D. and Castillo, J. C. (2018). Measurement model and invariance testing of scales measuring egalitarian values in ICCS 2009. En A. Sandoval-Hernandez, M. M. Isac and D. Miranda (Eds.) *Teaching Tolerance in a Globalized World*. Cham: Springer International Publishing

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika 58*(4), 525–43

Mullis, I.V.S., Martin, M.O., Foy, P., and Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center / IEA.

OECD. (2011). *Against the Odds: Disadvantaged Students who Succeed in School*. Paris: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing.

OECD. (2016a). *PISA 2015 Assessment and Analytical Framework.* Paris: OECD Publishing.

OECD. (2016b). PISA 2015 Background questionnaires. Annex A (pp. 129–196). En *PISA 2015 Assessment and Analytical Framework*. Paris: OECD Publishing.

Oliveri, M. E., and Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education*, *24*(4), 349–366. DOI: 10.1080/08957347.2011.607063

Oliveri, M. E., and von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333.

Oliveri, M. E., and von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*(1), 1–21. DOI: 10.1080/15305058.2013.825265

Rutkowski, L. and Rutkowski, D. (2010). Getting it "better": The importance of improving background questionnaires in International Large-Scale Assessment. *Journal of Curriculum Studies, 42*(3), 411–430. DOI: 10.1080/00220272.2010.487546

Rutkowski, L. and Rutkowski, D. (2017). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 0*(0), 1–14. DOI: 10.1080/00313831.2016.1261044

Rutkowski, L., Rutkowski, D. and Zhou, Y. (2016). Parameter estimation methods and the stability of achievement estimates and system rankings: Another look at the PISA model. *International Journal of Testing, 16*(1), 1–20.

Rutkowski, L., y Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31-57.

Schulz, W., Ainley, J. and Fraillon, J. (Eds.). (2011). *ICCS 2009 Technical Report*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Treviño, E., Fraser, P., Meyer, A., Morawietz, L., Inostroza, P. and Naranjo, E. (2015). *Informe de Resultados TERCE. Factores Asociados*. Santiago: Oficina Regional de Educación para América Latina y el Caribe (OREAL/UNESCO).

UNESCO-OREALC. (2016). *Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo, TERCE*. Santiago: Oficina Regional de Educación para América Latina y el Caribe (OREAL/UNESCO).

Van Der Linden, W. J. (2005). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement*, *42*(3), 283-302.

Received: 27-07-2017 Accepted: 14-09-2018

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Contact address: Andrés Sandoval-Hernández. University of Bath, Faculty of Humanities & Social Sciences, Department of Education. University of Bath, Claverton Down, Bath, BA1 6TP, United Kingdom. **E-mail:** A.Sandoval@bath.ac.uk

Received: 27-07-2017 Accepted: 14-09-2018