Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Corrigiendo las diferencias de uso de escala entre países de América Latina, Portugal y España en PISA

Correcting for Scale Usage Differences among Latin American Countries, Portugal, and Spain in PISA

He, Jia [(1)] & Van de Vijver, Fons [(2)]

(1) German Institute for International Educational Research. (2) Tilburg University.

Abstract

This paper investigated the effects of corrections for scale usage preference in seven Latin American countries, Portugal and Spain in student self-reports in the 2012 Programme for International Student Assessment (PISA). These targeted countries tend to show a strong tendency of expressing strong opinions and self-enhancement, which can pose serious validity threats in cross-cultural comparisons of self-reports. We examined to what extent score corrections, that have been proposed, would change the patterning of the cross-cultural differences. We corrected for the scale usage preferences in a measure of teacher support among 39,045 students in nine countries, based on extreme response style, overclaiming, and anchoring vignettes, respectively. These measures showed different effects: (1) All correction methods helped to improve measurement invariance, although the correction based on anchoring was less effective in reaching scalar invariance compared with the correction of extreme response style and overclaiming; (2) controlling for extreme response style and overclaiming changed the mean score of Spain to a greater extent than other countries, which seems to present a more realistic patterning, whereas the changes on correlations with other measures were rather limited. The use of anchored scores led to drastic changes both in means and correlations. A firm conclusion about which method is to be preferred cannot be given as there is no evidence which method enhances the validity of scores in these countries more. We discuss the necessity and practicability of correction methods.

Reception Date
2016 April 01

Approval Date
2016 June 22

Publication Date:
2016 June 23

Keywords:

Extreme response style; overclaiming; anchoring vignettes; comparability; validity; PISA

Resumen

En este trabajo se investigan los efectos de las correcciones en cuanto a la preferencia de uso de la escala en siete países de América Latina, Portugal y España mediante cuestionarios realizados a estudiantes del Programa para la Evaluación Internacional de Alumnos de 2012 (PISA). Estos países tienden a expresar opiniones de sólida autoestima, lo que puede representar graves amenazas de validez en las comparaciones transculturales de los cuestionarios. Se examinó cuánto podría cambiar esa puntuación de correcciones el patrón de las diferencias interculturales. Para las preferencias de uso de escala se corrigió en una medida de apoyo docente a un total de 39.045 estudiantes de 9 países mediante el estilo de respuesta extrema, sobrevaloración y viñetas de anclaje, respectivamente. Estas medidas mostraron diferentes efectos: (1) Todos los métodos de corrección ayudaron a mejorar la invariancia de medición, aunque, en comparación con la corrección de estilo de respuesta extrema y sobrevaloración, la corrección basada en el anclaje fue menos eficaz en alcanzar la invariancia escalar; (2) controlar el estilo de respuesta extrema y sobrevaloración cambia la puntuación media de España en mayor medida que en otros países, lo que parece presentar un patrón más realista, mientras que los cambios en las

Fecha de recepción
01 Abril 2016

Fecha de aprobación
22 Junio 2016

Fecha de publicación
23 Junio 2016

Autor de contacto / Corresponding author

He, Jia. Deutsches Institut für Internationale Pädagogische Forschung. Department of Educational Quality and Evaluation, Schloßstraße 29. 60486 Frankfurt am Main (Germany) jia.he@dipf.de

correlaciones con otras medidas fueron bastante limitados. El uso de las puntuaciones de anclaje llevó a cambios drásticos tanto en medios como en correlaciones. Dado que no se demuestra qué método proporciona más validez en sus puntuaciones, no se puede concluir de manera firme cuál es preferible. Se discute la necesidad y la viabilidad de los métodos de corrección.

Palabras clave:
Estilo de respuesta extrema; sobrevaloración; anclaje de viñetas; comparabilidad; validez; PISA.

En el Programa para la Evaluación Internacional de Alumnos (PISA) hay una famosa paradoja. A nivel individual, la correlación entre los autoinformes de actitud de Likert relacionados con rasgos positivos o ambiente de aprendizaje (p.ej. motivación y apoyo docente) y rendimiento tiende a ser positiva. No obstante, cuando las puntuaciones se suman a nivel nacional y la correlación se calcula entre los niveles medios de actitud y rendimiento de los países, encontramos una correlación negativa (He & Van de Vijver, 2015b; Kyllonen & Bertling, 2014). Es decir, los países de América Latina, que generalmente muestran puntuaciones por debajo de la media en los estudios del Informe PISA, tienden a obtener puntuaciones por encima de la media en autoinformes de actitud.

Esta paradoja implica desafíos en la comparabilidad de datos entre países. Cabe señalar que puede ser difícil comparar de forma completa todos los países del Informe PISA, dado el impacto de diversas culturas e idiosincrasias en las respuestas de los estudiantes a las medidas de la escala de Likert (p.ej. OECD, 2013b). Estamos interesados en las cuestiones de comparabilidad y validez de las respuestas de la escala de Likert en un grupo de países, concretamente los latinoamericanos, España y Portugal. Estos países comparten idiomas (español y portugués), y valores culturales (como se describe en el siguiente apartado), que podrían incidir en las preferencias de uso de la escala. Se han propuesto varios métodos para controlar las preferencias de uso de la escala (He & Van de Vijver, 2015b; Kyllonen & Bertling, 2014), incluyendo correcciones estadísticas y diseños innovadores de ítems; sin embargo, su eficacia para mejorar la comparabilidad y la validez de inferencias no se ha evaluado de manera sistemática. Por lo tanto, comparamos tres métodos para ajustar las diferencias de uso de escala y para discutir las implicaciones de los ajustes entre estos países del Informe PISA.

Preferencias de uso de la escala en países de América Latina, España y Portugal

La paradoja en los autoinformes de actitud y rendimiento académico en los estudios del Informe PISA puede estar influida por el uso diferencial de la escala por parte de los estudiantes de los diferentes grupos de países. Los países de América Latina, España y Portugal puntúan en un rango bastante alto en *evitación de la incertidumbre* y relativamente alto en *colectivismo* (Hofstede, 1980, 2009). La investigación ha demostrado que los encuestados en países con altos niveles de evitación de la incertidumbre tienden a ser intolerantes con la ambigüedad y por tanto apoyan más las categorías extremas en sus respuestas que las categorías medias (Harzing, 2006; He, van de Vijver, Domínguez, & Mui, 2014). Dentro de los países colectivistas, se hace una distinción más exhaustiva entre culturas del honor (p.ej. los países de la muestra) y culturas de modestia (p. ej. países de Asia Oriental). En países con una cultura del honor, los encuestados pueden defender su postura duramente y mostrar una mayor tendencia a mejorar la imagen personal (Smith, 2011; Uskul, Oyserman, & Schwarz, 2010). Incluso dentro de este grupo de países hay numerosas diferencias en el nivel de riqueza y en el contexto político e histórico, que posteriormente tienen un impacto en las preferencias de uso de escala y por tanto en la comparabilidad y en la validez de los autoinformes en la escala de Likert.

Métodos para controlar el uso diferencial de las escalas

Nos centramos en tres métodos que pueden aplicarse en los datos de estudiantes del Informe PISA en estos países para tener en cuenta las preferencias de uso de la escala: estilo de respuesta extrema, sobrevaloración y viñetas de anclaje.

El estilo de respuesta extrema consiste en que quienes responden tienden a usar mucho los puntos finales de una escala Likert de manera sistemática. (Paulhus, 1991). Chen, Lee, and Stevenson (1995) hallaron que los estudiantes de América Central y de América del Sur tendían a usar más el estilo de respuesta extrema que los de Asia del Este. Utilizando ítems de escala Likert para medir distintos constructos en el estudio del Informe PISA de 2012, extrajimos índices de estilo de respuesta extrema para cada estudiante y para sus países e investigamos el rol del estilo de respuesta extrema que prefería cada cultura en aspectos de comparabilidad y validez.

La sobrevaloración consiste en contestar pretendiendo que se conocen personas, eventos y productos que realmente no existen (Paulhus, Harms, Bruce, & Lysy, 2003): Se mide preguntando por el conocimiento del participante con respecto a una lista de conceptos tanto existentes como inexistentes. La sobrevaloración, medida por el número de señuelos que un participante dice conocer, es un indicador de la tendencia a la autoestima de los encuestados. Esta técnica se ha usado con los estudiantes encuestados en el Informe PISA de 2012 (OECD, 2013a). La técnica de sobrevaloración se ha desarrollado para identificar la tendencia a la autoestima, independiente de su propia habilidad.

Las viñetas de anclaje se refieren a proporcionar un punto de referencia común para los encuestados con distintas preferencias de uso de la escala (King, Murray, Salomon, & Tandon, 2004; King & Wand, 2007). Las viñetas son descripciones de personas imaginarias con distintos niveles del constructo de referencia. Los encuestados puntúan el nivel de rasgo de estas personas imaginarias con las mismas opciones de respuesta que la autoevaluación que se les pide que completen después de las viñetas. Se ajusta el sesgo de medida debido a las preferencias de uso de la escala de autoevaluación para producir una estimación del nivel real del rasgo buscado. Hay dos suposiciones de trabajo en las viñetas de anclaje: consistencia de la respuesta (es decir, los participantes utilizan los mismos mecanismos para responder las preguntas de autoevaluación y a las de viñeta) y equivalencia de las viñetas (es decir, las viñetas se entienden de la misma forma por todos los encuestados). El ajuste de la autoevaluación puede basarse en varios modelos. Aquí discutimos un acercamiento no paramétrico que ha sido utilizado en el Informe PISA, donde se valoraron tres viñetas de bajo, medio y alto nivel de apoyo docente en la misma escala que los ítems autoevaluados de apoyo docente. Este acercamiento consiste en reescalar las preguntas de autoevaluación (llamadas y) sobre la base de las respuestas de las cuestiones ordenadas por viñetas como J (llamadas z_1 a z_j) a una variable única de autoevaluación, denotada como C en la ecuación que aparece más abajo (King & Wand, 2007). Las valoraciones del autoinforme se reescalan en comparación con las valoraciones de las viñetas con el orden natural de puntuación de las viñetas (como se muestra en las fórmulas C). En el caso de respuestas equiparadas o inconsistentes ordenadas por viñetas (p.ej., $z_1 = z_2 = y$, o $z_2 > y = z_1$), las respuestas del autoinforme pueden adoptar un vector de valores posibles en lugar de un valor escalar. Por ejemplo, si las comparaciones del autoinforme con dos viñetas z_1 (nivel menor de rasgo) y z_2 (nivel mayor de rasgo) muestra un patrón de $z_2 > y = z_1$, C puede tomar cualquier valor desde 2 hasta 5. Esta técnica se ha aplicado en el cuestionario del Informe PISA de 2012 (OECD, 2013a) a estudiantes.

$$C = \begin{cases} 1 & if\ y < z_1 \\ 2 & if\ y = z_1 \\ 3 & if\ z_1 < y < z_2 \\ \dots & \\ 2J+1 & if\ y > z_j \end{cases}$$

El presente estudio

El presente estudio utiliza datos del cuestionario de contexto de los estudiantes y del rendimiento en matemáticas de los estudiantes en países Latinoamericanos, España y Portugal del Informe PISA de 2012 para comprobar si corrigiendo las preferencias de uso con los tres métodos mencionados puede mejorar (1) la comparabilidad de las medidas de una escala, concretamente Apoyo Docente (2) qué impacto tienen estos métodos en los patrones medios, y (3) en la correlación entre apoyo docente y rendimiento.

Método

Participantes

El cuestionario de estudiantes del Informe PISA de 2012 evaluaron competencias de los alumnos de 15 años en lectura, matemáticas y ciencias (con énfasis en matemáticas) en 60 países y economías (OCDE, 2013a). Los estudiantes fueron seleccionados mediante un procedimiento de muestreo estratificado para representar los colegios y las poblaciones de estudiantes de 15 años de cada país y economía, y completaron un cuestionario de contexto y un subconjunto del test cognitivo de distintas combinaciones que duró dos horas. Hay cuatro formas de cuestionarios de contexto con cuestiones parcialmente diferentes, que fueron distribuidas a una submuestra de estudiantes. Utilizamos datos de la Forma C del cuestionario de contexto de los estudiantes y datos de rendimiento en matemáticas en nueve países (Argentina, Brasil, Chile, Colombia, España, México, Perú, Portugal y Uruguay)[1]. Los tamaños de las muestras aparecen en la Tabla 1.

Tabla 1 *Estadísticas de las muestras*

País	Tamaño de muestra	Porcentaje de varones
Argentina	2.006	48
Brasil	6.381	48
Chile	2.272	49
Colombia	3.014	48
España	8.437	50
México	11.274	48
Perú	1.992	48
Portugal	1.913	49
Uruguay	1.756	48
Total	39.045	48

Medidas

El *apoyo docente* fue medido con cuatro ítems, con opciones de respuesta que oscilaban entre 1 (*muy de acuerdo*) a 4 (*muy en desacuerdo*). Los valores de Alfa de Cronbach para esta escala oscilaban entre .72 y .82 con una media de .77 en los 9 países.

Las puntuaciones del *estilo de respuesta extrema* se extrajeron de 15 ítems seccionados aleatoriamente de las autoevaluaciones de los estudiantes sobre aprendizaje y docencia (excluyendo los ítems sobre apoyo docente) con 4 opciones de respuesta en el cuestionario de contexto de los estudiantes. El promedio de la correlación inter-ítem fue de .15, indicando una razonable heterogeneidad en los ítems para captar la tendencia de respuesta sistemática en lugar de un rasgo sustantivo. Las respuestas a estos ítems se recodificaron con respuestas de 1 y 4 como un 1, y los otros valores como un 0. La fiabilidad de los ítems recodificados oscilaron entre .57 a .69 en los diversos países con una media de .61. La media de los ítems recodificados fue tomada como un índice de estilo de respuesta extrema.

Se aplicaron tres ítems *sobrevalorados* (es decir, ítems referidos a conceptos que no existen) junto con ítems de familiaridad con

[1] Sólo en la forma C del cuestionario de contexto del alumno se han administrado las medidas elegidas (apoyo docente, sobrevaloración, y anclaje de viñetas sobre apoyo docente). Esta submuestra permite estimar medias menos sesgadas.

conceptos matemáticos. Las opciones de respuesta oscilaban entre 1 (*nunca lo he oído*) a 5 (*lo conozco bien, entiendo el concepto*), y la fiabilidad osciló entre .47 y .75 entre países, con una media de .67. La puntación media de los tres ítems se tomó como una puntuación de sobrevaloración.

Se aplicó un conjunto de *viñetas de anclaje* con viñetas de apoyo docente a los deberes bajo, medio y alto al reescalado de apoyo docente. Las opciones de respuesta fueron las mismas que los ítems de escala de apoyo docente. El reescalado de los ítems de apoyo docente se realizó en el paquete de anclajes en R, utilizando el enfoque no paramétrico (Wand & King, 2007). En los casos de empate y desordenaciones, las respuestas graduadas tenían una gama de valores posibles y la más alta calificación posible fue utilizada como representación. La escala de anclaje de apoyo docente obtuvo un rango en fiabilidad desde .88 hasta.92. con una media de .90.

El autoinforme de los estudiantes de *instrucción dirigida por el profesor* comprendía cinco ítems con una escala de 4 puntos desde 1 (*en todas las clases*) a 4 (*nunca o casi nunca*). La puntuación final fue codificada al revés, por lo que una puntuación elevada indicaba que el docente dirigía mucho. La fiabilidad osciló entre .67 y .75 con una media de .70.

El rendimiento de los estudiantes en matemáticas fue medido con diferentes subgrupos del test cognitivo y estimado con valores plausibles. Los valores plausibles son valores imputados que se parecen a las puntuaciones individuales en el test y que tienen aproximadamente la misma distribución que el rasgo latente medido. Se produjeron cinco valores plausibles de rendimiento en matemáticas para cada estudiante.

Resultados

Describimos los resultados en tres partes: la prueba de invarianza de medida de la escala de apoyo docente, las pruebas de las diferencias de medias y las asociaciones de apoyo docente y de instrucción dirigida por el profesor, y el rendimiento en matemáticas de los estudiantes con y sin correcciones.

Las pruebas de invarianza de medida

Probamos la invarianza de medida de apoyo docente en cuatro casos: (1) con puntuaciones directas, (2) con estilo de respuesta extrema corregida (es decir, el estilo de respuesta extrema observado prediciendo cada ítem de respuesta observado, y la predicción de la respuesta a los cuatro ítems por el factor latente del apoyo docente); (3) con sobrevaloración corregida (es decir, igual que el segundo caso); (4) con puntuaciones de los ítems ajustadas por el anclaje. La prueba de invarianza de medida fue realizada utilizando el análisis factorial confirmatorio multigrupo con AMOS (Arbuckle, 2006). Se comprobaron tres niveles de invarianza: invarianza configuracional (es decir, el constructo se mide con los mismos ítems entre países), invarianza métrica (es decir, forzar que las cargas factoriales sean las mismas entre países) y la invarianza escalar (es decir, se forzó tanto las cargas factoriales como las intersecciones de los ítems para que fuesen iguales entre los países). Con la invarianza métrica se pueden comparar las asociaciones entre variables en cada país, mientras que sólo con la invarianza escalar se pueden comparar directamente las puntuaciones escalares entre países (van de Vijver & Leung, 1997). El ajuste al modelo fue evaluado mediante pruebas de chi-cuadrado, el Índice de Ajuste Comparativo (CFI) (aceptable por encima de .90), y el Error Cuadrático Medio de Aproximación (RMSEA, aceptable inferior a .06); la aceptación de un modelo más restrictivo basado en el cambio del valor del CFI de menos de .01 desde el modelo menos restringido hasta el más restringido (Cheung & Rensvold, 2002).

Los índices de ajuste del modelo para todos los modelos se presentan en la tabla 2. En todos los casos, se alcanza la invarianza configuracional y métrica. La invarianza escalar se alcanzó sólo cuando se controlaron el estilo de respuesta extrema y la sobrevaloración.

Tabla 2 *Modelo de ajuste de las pruebas de medida de invarianza*

	χ^2	*df*	CFI	RMSEA	ΔCFI	ΔRMSEA
Puntuaciones Directas						
Configuracional	53.962**	18	.999	.007		
Métrica	331.735**	42	.993	.013	-.006	.006
Escalar	2520.93**	74	.941	.029	-.052	.016
Corrección por Estilo de Respuesta Extrema						
Configuracional	53.526**	18	.999	.007		
Métrica	325.068**	42	.993	.013	-.006	.006
Scalar	486.821**	74	.990	.012	-.003	-.001
Corrección por sobrevaloración						
Configuracional	53.611**	18	.999	.007		
Métrica	328.073**	42	.993	.013	-.006	.006
Escalar	636.253**	74	.987	.014	-.006	.001
Puntuaciones ancladas						
Configuracional	91.073**	18	.999	.010		
Métrica	267.398**	42	.997	.012	-.002	.002
Escalar	2601.131**	74	.972	.030	-.025	.018

** $p < .01$.

A pesar de que las puntuaciones de anclaje no alcanzaron la invarianza escalar, el ajuste del modelo de las puntuaciones de anclaje fue mejor que los de las puntuaciones directas, indicados por el cambio del valor de CFI del modelo de la invarianza métrica al escalar de .25 y .52 en estos dos casos respectivamente. Resumiendo, controlando el estilo de respuesta extrema o la sobrevaloración aumentó la comparabilidad de las puntuaciones en estos nueve países. El anclaje de viñetas mejoró la comparabilidad en cierta medida, aunque no llegó a la completa comparabilidad.

Los patrones de medias antes y después de la corrección

Las puntuaciones medias latentes de la escala de apoyo docente para cada país fueron calculadas mediante el Análisis Factorial Confirmatorio Multigrupo en los cuatro casos. México fue tratado como el grupo de referencia, debido al tamaño de muestra mayor en este país. A nivel técnico, la media latente de México se forzó a ser cero en el modelo de invarianza escalar, y las medias latentes de los otros países se estimaron libremente. La comparación de los patrones de medias con el 95 % de intervalos de confianza está dibujada en la Figura 1. Conviene observar que las puntuaciones no están recodificadas a la inversa; por tanto, las medias en la Figura 1 representan el nivel de falta de apoyo docente.

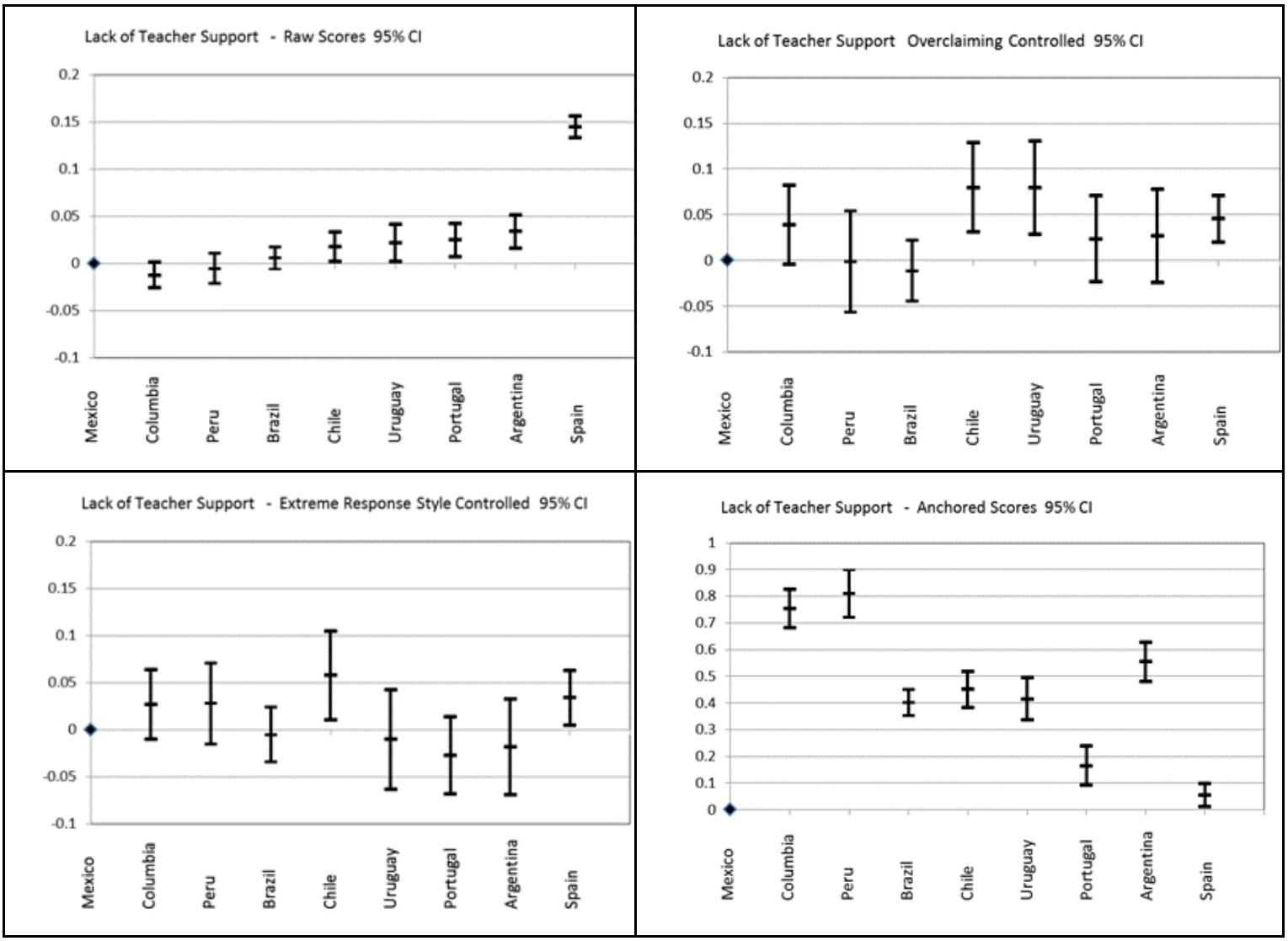

Figura 1. *Los patrones de medias en puntuaciones directas y ajustadas con falta de apoyo docente (México como país de referencia)*

Con las puntuaciones directas, muchos de los países latinoamericanos no difieren mucho en este constructo, excepto España, que mostró un menor nivel de apoyo docente en comparación con los otros países. Cuando se corrigió el estilo de respuesta extrema, el patrón de media cambió en dos formas notables. En primer lugar, aumentaron todos los intervalos de confianza, indicando que se debería tener más en cuenta los errores de medida. En segundo lugar, la diferencia entre España y otros países era mucho menor debido a la corrección. Se observó un patrón similar cuando se controló la sobrevaloración. En todos los casos mencionados, las diferencias en apoyo docente entre países fueron bastante limitadas, mientras que el patrón cambió drásticamente cuando se usaron las puntuaciones de anclaje. Con las puntuaciones de anclaje, Colombia y Perú mostraron un nivel significativamente menor de apoyo docente en comparación con España, Portugal y México. En otras palabras, los efectos de corrección sobre patrones de medias fueron algo diferentes, con el enfoque de anclaje de viñetas mostrando el mayor cambio y los otros dos enfoques un cambio más limitado.

Correlaciones antes y después de la corrección

Se esperaba que apoyo docente podría correlacionar positivamente con la enseñanza

dirigida por el profesor. Las expectativas sobre la correlación entre apoyo docente y logro académico están menos claras. Por una parte, estas escalas deberían estar correlacionadas en sentido positivo, debido a que las interacciones positivas con los profesores contribuyen a un mejor aprendizaje. Por otra parte, los estudiantes que recibieron mayor apoyo docente podrían ser los que tenían buenos resultados, por lo que no es irrazonable una correlación negativa. En la Tabla 3 aparecen las correlaciones con las puntuaciones factoriales directas y factoriales de anclaje de apoyo docente, las correlaciones parciales con el estilo de respuesta extrema y la sobrevaloración controlada.

Tabla 3 *Correlaciones de apoyo docente en cada país*

	Correlación con Instrucción Directa del Profesor				Correlación con rendimiento en Matemáticas (PV1)			
	Directas	Ajustada por ERS	Ajustada por Sobrevaloración	Anclada	Directas	Ajustada por ERS	Ajustada por Sobrevaloración	Anclada
Argentina	.599	.561	.601	.121	-.044	*-.041*	-.073	.134
Brasil	.592	.561	.592	.154	-.038	*-.006*	-.049	.150
Chile	.654	.617	.641	.230	-.045	-.066	*-.042*	.187
Colombia	.592	.522	.599	.163	.044	*.026*	*.032*	.182
España	.648	.642	.653	.264	-.027	*-.029*	-.028	.136
México	.612	.567	.608	.150	-.028	-.056	-.045	.146
Perú	.627	.541	.621	.135	-.030	*-.049*	-.062	.138
Portugal	.662	.632	.656	.235	-.040	-.049	-.036	.143
Uruguay	.597	.582	.592	.165	-.101	-.099	-.106	.126

Nota., Las correlaciones con el rendimiento en matemáticas se realizaron con los cinco valores plausibles, y todos mostraron los mismos resultados, por tanto, sólo se presentan la correlaciones con el primer valor plausible, (PV1). ERS = estilo de respuesta extrema. Todas las correlaciones son significativas a p <0,01 excepto las cursiva

En los cuatro casos, el apoyo docente y la enseñanza dirigida por el profesor están relacionadas positivamente. La corrección basada en el estilo de respuesta extrema y la sobrevaloración tuvo un efecto más bien limitado; la ligera reducción en las correlaciones indica que algunas preferencias de uso de la escala general estaban parcialmente eliminadas. El cambio en el tamaño de la correlación fue más evidente en las puntuaciones de anclaje. Las correlaciones de rendimiento en matemáticas en puntuaciones directas, la corrección de estilo de respuesta extrema y la corrección de sobrevaloración en general fueron ligeramente negativas, mientras que con las puntuaciones de anclaje, el apoyo docente correlacionaba positivamente con el logro en matemáticas. Esto apunta a la eficacia de las viñetas de anclaje en revertir las correlaciones entre experiencia positiva en el aprendizaje y logro académico. Sin embargo, todavía no está claro si las puntuaciones de anclaje son más válidas.

Discusión

Estudiamos los efectos de tres métodos de corrección de las escalas Likert de autoinformes en nueve países con preferencias de uso de la escala similares según los datos del Informe PISA de 2012. Estos países (Latinoamericanos y de Europa del Sur) generalmente muestran un fuerte tendencia en expresividad y en autoestima, que pueden influir en la validez de los autoinformes. Examinamos el impacto de las correcciones para el estilo de respuesta extrema, sobrevaloración y viñetas de anclaje sobre la invarianza de medida, la media del país y la correlación con variables externas. Los principales resultados incluyen: (1) Todas las correcciones ayudan a mejorar el nivel de invarianza de medida, aunque las puntuaciones

de anclaje fueron menos efectivas en alcanzar la invarianza de la escala en comparación con la corrección del estilo de respuesta extrema y la sobrevaloración; (2) Controlar el estilo de respuesta extrema y la sobrevaloración tiene efectos limitados sobre las puntuaciones medias de país o la correlación con otras variables, mientras que las puntuaciones de anclaje mostraron cambios más drásticos. No hay evidencia de que la corrección mejora la validez de las puntuaciones en estos países. La mejora en las estadísticas de invarianza que utilizan principalmente las correcciones del estilo de respuesta extrema y la sobrevaloración indican que las correcciones de las puntuaciones pueden mejorar la validez. Además, el gráfico de medias del país después de las correcciones de estos sesgos es más atractivo e intuitivo en el que la media de España ahora se acerca más a la media de los otros países. Sin embargo, la falta de cualquier impacto de estas correcciones sobre las correlaciones es contraintuitivo. Una gran puntuación en la medida del estilo de respuesta extrema global podría estar también presente en la puntuación de apoyo docente (la correlación de estos dos es .25), lo que representa una considerable reducción de la puntuación. Estas correlaciones están más afectadas por el anclaje. Sin embargo, no hay evidencia de que estas correlaciones sean más reales (ni para enseñanza dirigida por el profesor ni para el rendimiento en matemáticas).

Se ciernen grandes amenazas potenciales a la validez de las preferencias diferenciales de uso de la escala en los países latinoamericanos, especialmente en contextos de evaluaciones a gran escala, donde los datos comparativos entre culturas se usan para influir en las decisiones políticas basadas en la evidencia (p.ej. Goldstein, 2004; Gorur, 2014). En el Informe PISA la paradoja inversa de las correlaciones a nivel individual y nacional entre las experiencias positivas de autoinformes y rendimiento crearon la necesidad de corregir las preferencias de uso de la escala. No obstante, en el presente estudio, la corrección con distintos métodos en siete países latinoamericanos y dos países de Europa del Sur mostró resultados mixtos: la corrección con el estilo de respuesta extremo y la sobrevaloración aseguró la total invarianza de la escala en la escala de apoyo docente, aunque no cambió los patrones de correlación con variables externas. Las viñetas de anclaje cambiaron los patrones de correlación pero no ayudaron a alcanzar la invarianza de la escala. Parece que estos métodos de corrección apuntan a diferentes preferencias de uso de la escala. En la escala de uso por lo general se prefiere el estilo de respuesta extrema y la sobrevaloración, lo que afecta uniformemente a todo tipo de autoinformes. Las viñetas de anclaje se centran en las diferencias individuales de manera más específica en la interpretación del contenido y en las opciones de respuesta. El escalamiento basado en las viñetas de anclaje puede conseguir una mayor variación en las puntuaciones tanto a nivel individual como cultural. Sin embargo, no está claro si el cambio de escala presenta otro tipo de sesgo, en particular, dada la probable violación de las suposiciones (estrictas) de las viñetas de anclaje. Se puede concluir que las correcciones de las puntuaciones tales como las derivadas de las viñetas de anclaje son capaces de revertir la paradoja del rendimiento de la motivación cuando se comparan diferentes regiones, como Asia del Este y Latinoamérica, estos procedimientos producen resultados que son mucho más difíciles de interpretar cuando se aplica en una región con una cultura más homogénea.

Existen diferentes procedimientos que presumiblemente pueden mejorar la comparabilidad y la validez de los datos de autoevaluación entre países. Es difícil determinar hasta qué punto se puntúa por encima el constructo sustantivo que se mide, porque las preferencias del uso de la escala podrían ser una parte integral de la estructura psicológica de los participantes (He & van de Vijver, 2015c), y se suma al nivel cultural, pueden representar aspectos importantes de la cultura nacional, como las preferencias de valores individualistas o colectivistas (Smith, 2004, 2011). Algunos estudios demuestran

efectos de corrección significativos en las comparaciones entre países (p.ej., Diamantopoulos, Raeynolds, & Simintiras, 2006), mientras que otros estudios informan sobre efectos insignificantes (p.ej., He & van de Vijver, 2015a). Los resultados de nuestro estudio muestran resultados inconsistentes entre los métodos de corrección. Para saber qué método de corrección reduce el sesgo y aumenta la validez de diferentes medidas se requiere más investigación.

En un sentido práctico, todavía vale la pena el esfuerzo de comparar las puntuaciones con y sin varias correcciones. Tanto el estilo de respuesta extrema como la sobrevaloración funcionan de manera semejante en cuanto a sus efectos de corrección, y con la consideración de que el estilo de respuesta extrema se puede construir con diferentes respuestas con ítems existentes, mientras que la sobrevaloración requiere una medida adicional; parece que la corrección para el estilo de respuesta extrema es más fácil de implementar. El uso de las viñetas de anclaje requiere precaución; no se puede informar de resultados concluyentes con puntuaciones de anclaje hasta que los dos supuestos de este método estén empíricamente probados y verificados.

Limitaciones y líneas futuras

Este estudio presenta algunas limitaciones. En primer lugar, sólo consideramos nueve de los 64 países del Informe PISA. Esto se debe a que las tendencias de autoestima y expresividad en estos países son especialmente preocupantes en la sobrevaloración de los autoinformes. Nuevos estudios podrán abordar la investigación en contextos culturales más variados. En segundo lugar, restringimos nuestro análisis a los estudiantes que contestaron la Forma C del cuestionario para evitar un número excesivo de valores perdidos. Por tanto, las medias estimadas de los países están basadas en aproximadamente un tercio de la muestra total, que puede no ser totalmente representativa a nivel nacional. Finalmente, limitamos nuestros métodos de corrección debido a la disponibilidad de los datos. Hay otros métodos de diseño de ítems como las preguntas de elección forzada y las preguntas de juicio situacional y otras correcciones estadísticas como los modelos bifactor que pueden ayudar a remediar la falta de validez y comparabilidad (e.g., Brown & Maydeu-Olivares, 2011; Cheung & Rensvold, 2000; Rutkowski et al., 2014). Consideramos que cuantos más esfuerzos se pongan en reducir el sesgo de medición en distintos contextos culturales, se pueden utilizar mejor los datos de evaluación a gran escala para la investigación básica y la formulación de políticas basadas en evidencias.

Referencias

Arbuckle, J. L. (2006). *AMOS user's guide*. Chicago, IL: SPSS.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*, 460-502. http://dx.doi.org/10.1177/0013164410375112

Chen, C., Lee, S.-y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*, 170-175. doi: http://dx.doi.org/10.1111/j.1467-9280.1995.tb00327.x

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*, 187-212. doi: http://dx.doi.org/10.1177/0022022100031002003

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255. doi: http://dx.doi.org/10.1207/s15328007sem0902_5

Diamantopoulos, A., Raeynolds, N. L., & Simintiras, A. C. (2006). The impact of response styles on the stability of cross-

national comparisons. *Journal of Business Research, 59*, 925-935. doi: http://dx.doi.org/10.1016/j.jbusres.2006.03.001

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

Gorur, R. (2014). Towards a sociology of measurement in education policy. *European Educational Research Journal, 13*, 58-72. doi: http://dx.doi.org/10.2304/eerj.2014.13.1.58

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management, 6*, 243-266. doi: http://dx.doi.org/10.1177/1470595806066332

He, J., van de Vijver, F., J. R., Domínguez, A. d. C., & Mui, P. H. C. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross-Cultural Management, 14*, 306-322. doi: http://dx.doi.org/10.1177/1470595814541424

He, J., & van de Vijver, F. J. R. (2015a). Effects of a general response style on cross-cultural comparisons: Evidence from the Teaching and Learning International Survey. *Public Opinion Quarterly, 79*, 267-290. doi: http://dx.doi.org/10.1093/poq/nfv006

He, J., & Van de Vijver, F. J. R. (2015b). The motivation-achievement paradox in international educational achievement tests: Toward a better understanding. In R. B. King & A. B. I. Bernardo (Eds.), *The psychology of Asian learners: A festschrift in honor of David Watkins* (pp. 253-268). Singapore: Springer.

He, J., & van de Vijver, F. J. R. (2015c). Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences, 31*, 129-134. doi: http://dx.doi.org/10.1016/j.paid.2014.09.009

Hofstede, G. (1980). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Beverly Hills, CA: Sage.

Hofstede, G. (2009). *Dimension data matrix*. Dimension Data Matrix Retrieved 03/02/2001 http://www.geerthofstede.eu/dimension-data-matrix

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*, 191-207. doi: http://dx.doi.org/10.1017/S00030554040010 8X

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis, 15*, 46-66. doi: http://dx.doi.org/10.1093/pan/mpl011

Kyllonen, P. C., & Bertling, J. P. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. v. Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277-286). Boca Raton, FL: CRC Press.

OECD. (2013a). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD Publishing.

OECD. (2013b). *PISA 2012 technical report*. Paris, France: OECD Publishing.

Paulhus, D. L. (1991). Measurement and control of response biases. In J. Robinson, P. Shaver & L. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17-59). San Diego, CA: Academic Press.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology, 84*, 890-904. doi: http://dx.doi.org/10.1037/0022-3514.84.4.890

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: CRC Press.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35*, 50-61. doi: http://dx.doi.org/10.1177/0022022103260380

Smith, P. B. (2011). Communication styles as dimensions of national culture. *Journal of Cross-Cultural Psychology, 42*, 216-233. doi: http://dx.doi.org/10.1177/0022022110396866

Uskul, A. K., Oyserman, D., & Schwarz, N. (2010). Cultural emphasis on honor, modesty or self-enhancement: Implications for the survey response process. In J. A. Harkness, M. Broun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multiregional and multicultural contexts* (pp. 191-201). New York, NY: Wiley.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.

Wand, J., & King, G. (2007). *Anchoring vignetttes in R: A (different kind of) vignette*. Retrieved from http://wand.stanford.edu/anchors/doc/anchors.pdf

Author / Autor

To know more / Saber más

He, Jia (jia.he@dipf.de).

Doctora en Psicología Transcultural por la Universidad de Tilburg (Holanda), 2011-2015. Master en Comunicación Intercultural, Universidad de Estudios Internacionales de Shanghai (China). Licenciada Marketing, Universidad de Dongbei de Finanzas y Economía (Dalian, China). Post Doctorado en el Deutsches Institut für Internationale Pädagogische Forschung (DIPF) desde 2015. Investiga sobre la comparabilidad de los datos de autoinforme en los estudios internacionales a gran escala y los nuevos formatos de ítems, entre otras cosas. Consultora de Investigación del Banco Mundial desde 2015. Analista OCDE desde 2015. Dirección postal: Department of Educational Quality and Evaluation, Schloßstraße 29. 60486 Frankfurt am Main (Alemania).

Van de Vijver, Fons (fons.vandevijver@tilburguniversity.edu)

Profesor de Psicología Transcultural en la Universidad de Tilburg, en la Universidad del Noroeste (Sudáfrica) y en la Universidad de Queensland (Australia). Especializado en la investigación transcultural y en "métodos y análisis de datos de la investigación comparativa". Van de Vijver recibió su MA en 1991 y su doctorado en Psicología en la Universidad de Tilburg. En 2013 recibió junto con María Cristina Richaud el premio de la APA por sus contribuciones distinguidas a la Promoción Internacional de Psicología.

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

[ISSN: 1134-4032]