Correcting for Scale Usage Differences among Latin American Countries, Portugal, and Spain in PISA (Corrigiendo las diferencias de uso de escala entre países de América Latina, Portugal y España en PISA)

He, Jia; van de Vijver, Fons

Published in:
Revista ELectrónica de Investigación y EValuación Educativa

Publication date:
2016

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

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

Corrigiendo las diferencias de uso de escala entre países de América Latina, Portugal y España en 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 investigaron los efectos de las correcciones sobre la preferencia de uso de la escala en siete países de América Latina, Portugal y España en cuestionarios de estudiantes en el Programa para la Evaluación de Estudiantes 2012 (PISA). Estos países destinatarios tienden a mostrar una tendencia de expresar opiniones fuertes y de auto-mejora, lo que puede plantear amenazas graves de validez de las comparaciones transculturales de los cuestionarios. Se examinó en qué medida la puntuación de correcciones, que se han propuesto, podría cambiar el patrón de las diferencias culturales. Hemos corregido para las preferencias de uso de la escala de una medida de ayuda al profesor de entre 39,045 estudiantes en nueve países, con base en el tipo de respuesta extrema, overclaiming, y el anclaje de viñetas, respectivamente. Estas medidas mostraron diferentes efectos: (1) Todos los métodos de corrección ayudaron a mejorar la invariancia de medición, a pesar de que la corrección sobre la base de anclaje fue menos eficaz en alcanzar la invariancia escalar en comparación con la corrección de estilo de respuesta extrema y overclaiming; (2) el control de estilo de respuesta extrema y overclaiming cambia la puntuación media de España en mayor medida que en otros países, lo que parece

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

presentar un patrón más realista, mientras que los cambios en las correlaciones con otras medidas fue bastante limitado. El uso de las puntuaciones de anclaje llevó a cambios drásticos tanto en medios como en correlaciones. Una conclusión firme sobre qué método es preferible, no puede ser ofrecido ya que no hay evidencia de que el método mejore la validez de las puntuaciones en estos países. Se discute la necesidad y la viabilidad de los métodos de corrección.

Palabras clave:

Estilo de respuesta extrema; sobreestimación; anclaje de viñetas; comparabilidad; validez; PISA.

There is a famous paradox in the Programme for International Student Assessment (PISA). At individual level, the correlation between Likert self-report attitudes related to positive traits or learning environment (e.g., motivation and teacher support) and achievement tends to be positive. However, when scores are aggregated at country level and the correlation is computed between countries' average levels of attitude and achievement, a negative correlation is found (He & Van de Vijver, 2015b; Kyllonen & Bertling, 2014). That is, Latin American countries, typically showing lower than average scores on achievement in the PISA studies, tend to have higher than average scores on self-report attitudes.

Such a paradox suggests challenges in the comparability of data across countries. It is noted that full comparability of all PISA countries might be hard to achieve, given the impact of diverse cultures on and idiosyncrasies in students' responses to Likert-scale measures (e.g., OECD, 2013b). We are interested in the comparability and validity issues of Likert-scale responses in a cluster of countries, namely Latin American countries, Spain, and Portugal. These countries share languages (i.e., Spanish and Portuguese), and they share cultural values (as described in the next section), which might have a bearing on the scale usage preferences. Various methods including statistical corrections and innovative item designs have been proposed to control for scale usage preferences (Rutkowski, von Davier, & Rutkowski, 2014), yet their effectiveness for improving comparability and validity of inferences has not been systematically evaluated. Therefore, we compare three methods to adjust scale usage differences and discuss the implications for adjustments among these PISA countries.

Scale Usage Preferences in Latin American Countries, Spain and Portugal

The paradox in self-reported attitudes and academic achievement in the PISA studies may be affected by differential scale usage by students in different cluster of countries. Latin American countries, Spain and Portugal rank rather high on uncertainty avoidance and relatively high on collectivism (Hofstede, 1980, 2009). Research has shown that survey respondents in countries with high levels of uncertainty avoidance tend to be intolerant of ambiguity, and thus endorse more extreme categories in their responses than middle categories (Harzing, 2006; He, van de Vijver, Domínguez, & Mui, 2014). Within collectivistic countries, a finer distinction is made between honor cultures (e.g., our target countries) and modesty cultures (e.g., East-Asian countries). In countries with an honor culture, survey respondents may defend their positive fiercely, and may show a higher tendency to enhance the personal image (Smith, 2011; Uskul, Oyserman, & Schwarz, 2010). Even within this cluster of countries, there are numerous differences in affluence level, political and historical background, which subsequently impact on scale usage preferences and further impact on the comparability and validity of Likert-scale self-reports.

Methods to Control for Scale Usage Differences

We target three methods that can be applied in the PISA student data in these countries to account for the scale usage preferences:

extreme response style, overclaiming, and anchoring vignettes.

Extreme response style refers to the systematic tendency of respondents to over-use the endpoints of a Likert scale (Paulhus, 1991). Chen, Lee, and Stevenson (1995) found that Central and South American students are more inclined to use extreme response style than East Asians. Using Likert items measuring various constructs in the 2012 PISA study, we extract indexes of extreme response styles for each student and their countries and investigate the role of culturally preferred extreme response style on comparability and validity issues.

Overclaiming refers to claiming to have knowledge of nonexistent persons, events, and products, responses (Paulhus, Harms, Bruce, & Lysy, 2003). It is measured by asking for participant's knowledge of concepts in a list of existing and non-existing concepts. Overclaiming, measured by the number of foils a participant claims to know, is an indicator of respondents' self-enhancement tendency. This technique has been used in the PISA student questionnaire in 2012 (OECD, 2013a). The overclaiming technique is developed to capture the self-enhancement tendency independent of one's ability.

Anchoring vignettes involve an approach to provide a common reference point for respondents with different scale usage preferences (King, Murray, Salomon, & Tandon, 2004; King & Wand, 2007). Vignettes are descriptions of hypothetical persons with different levels of the target trait. Respondents rate the trait level of these hypothetical persons on the same response options as the self-assessment that they are requested to fill out after the vignettes. The measurement bias due to scale usage preferences from the self-assessment is adjusted to yield an estimate of the actual level of the target trait. There are two working assumptions of anchoring vignettes: response consistency (i.e., participants use the same mechanisms to give responses to self-assessment questions and the vignette questions) and vignette equivalence (i.e., the vignettes are understood by all respondents in the same way). The adjustment of the self-assessment can be based on various models. We discuss here a nonparametric approach that has been used in PISA, where three vignettes of low, medium and high trait levels of teacher support were rated on the same scale as the self-reported teacher support items. This approach is to rescale self-assessment responses (denoted as y) on the basis of responses of J ordered vignette questions (denoted as z1 to zj) to a single variable self-assessment, denoted by C in the equation below (King & Wand, 2007). With natural ordering of rating on vignettes, self-report rate is rescaled in comparisons to the vignette rating (as shown in the formulas C). In case of tied or inconsistently ordered vignette responses (e.g., z1 = z2 = y, or z2 > y = z1), the self-assessment responses can take a vector of possible values instead of one scalar value. For instance, if the comparisons of self-assessment y with two vignettes z1 (lower trait level) and z2 (higher trait level) shows a pattern of z2 > y = z1, C may take any of the values from 2 to 5. This technique has been applied in the PISA 2012 student questionnaire (OECD, 2013a).

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

The Present Study

The present study makes use of data of the student background questionnaire and students' math achievement in Latin American countries, Spain, and Portugal in the 2012 PISA to check whether correcting scale usage preferences with the three above mentioned methods can improve (1) the measurement comparability of one target scale, namely Teacher Support, (2) what impact these methods make on the mean patterns, and (3) on the correlation between teacher support and achievement.

Method

Participants

The PISA student survey in 2012 assessed competencies of 15-year-olds in reading, mathematics, and science (with a focus on mathematics) in over 60 countries and economies (OECD, 2013). Students were recruited through a stratified sampling procedure to represent the schools and the 15-year-old student populations of each country and economy, and they took a background questionnaire and a subset of the cognitive test of different combinations that lasted two hours. There are four forms of student background questionnaires with partially different questions, which were distributed to a subsample of students. We used data on the Form C student background questionnaire and the math achievement data in nine countries (Argentina, Brazil, Chile, Colombia, Spain, Mexico, Peru, Portugal, and Uruguay)[1] . Sample sizes per country are presented in Table 1.

Table 1 *Sample Statistics*

Country	Sample Size	Percentage of Males
Argentina	2,006	48
Brazil	6,381	48
Chile	2,272	49
Colombia	3,014	48
Spain	8,437	50
Mexico	11,274	48
Peru	1,992	48
Portugal	1,913	49
Uruguay	1,756	48
Total	39,045	48

Measures

Teacher Support was measured with four items, with response options ranging from 1 (*strongly agree*) to 4 (*strongly disagree*). Values of Cronbach's Alpha for this scale ranged from .72 to .82 with a mean of .77 in the 9 countries.

Extreme response style scores were extracted from 15 randomly selected items on student self-reports of learning and teaching (excluding items on teacher support) with 4-point response options in the student background questionnaire. The average inter-item correlation was .15, indicating reasonable item heterogeneity to capture the systematic response tendency rather than a substantive trait. The responses on these items were recoded with responses of 1 and 4 as 1, and other values as 0. The reliability of the recoded items ranged from .57 to .69 across countries with a mean of .61. The mean of the recoded items was taken as an index of extreme response style.

Three *overclaiming* items (i.e., items referring to concepts that do not exist) were administered along with items on the familiarity with math concepts. The response option ranged from 1 (*never heard of it*) to 5 (*know it well, understand the concept*), and reliability ranged from .47 to .75 across countries, with a mean of .64. The mean rating of the three items was taken as an overclaiming score.

A set of *Anchoring Vignettes* with vignettes about low, medium, and high teacher support on homework were applied to the rescaling of teacher support. The response options were the same as the teacher support scale items. The rescaling of teacher support items were carried out in the anchors package in R, using the nonparametric approach (Wand & Kimg, 2007). In cases of ties and misorderings, the rescaled responses had a range of possible values, and the highest possible rating was used as a proxy. The anchored scale of teacher support had a reliability ranging from .88 to .92 with a mean of .90.

Students' self-report *Teacher-Directed Instruction* comprised five items answered on a 4-point scale from 1 (*Every Lesson*) to 4 (*Never or Hardly Ever*). The final scale score was reverse coded, so a higher score indicated higher teacher-directed instruction. The reliability ranged from .67 to .75 with a mean of .70.

Students' *math achievement* was measured with different subsets of the cognitive test and was estimated using plausible values. Plausible values are imputed values that resemble individual test scores and have approximately the same distribution as the latent trait being measured. Five plausible values of math achievement for each student were produced.

Results

We describe the results in three parts: the measurement invariance test of the teacher support scale, the tests of the mean differences, and the associations of teacher support and teacher-directed instruction and student math achievement with and without corrections.

The Measurement Invariance Tests

We tested the measurement invariance of teacher support in four cases: (1) with raw scores; (2) with extreme response style corrected for (i.e., the observed extreme response style predicting each observed item response, and all the four observed item response predicted by the latent factor of teacher support); (3) with overclaiming corrected for (i.e., same as the second case); (4) with anchoring-adjusted item scores. The measurement invariance test was performed using multigroup confirmatory factor analysis in AMOS (Arbuckle, 2006). Three levels of invariance were checked: configural invariance (i.e., the construct is measured by the same items across countries), metric invariance (i.e., factor loadings were constrained to be equal across countries), and scalar invariance (i.e., both factor loadings and item intercepts were both constrained to be equal across countries). With metric invariance, associations between variables in each country can be compared, whereas only with scalar invariance can scale scores be directly compared across countries (van de Vijver & Leung, 1997). The model fit was evaluated by Chi-square tests, Comparative Fit Index (acceptable above .90), and Root Mean Square Error of Approximation (acceptable below .06); the acceptance of a more restricted model was based on change of CFI value of less than .01 from the less to the more restricted model (Cheung & Rensvold, 2002).

The model fit indexes for all models are presented in Table 2. In all cases, configural and metric invariance were achieved. Scalar invariance was only achieved when extreme response style and overclaiming were controlled for.

Table 2 *Model Fit of Measurement Invariance Tests*

	χ^2	*df*	CFI	RMSEA	ΔCFI	ΔRMSEA
Raw Scores						
Configural	53.962**	18	.999	.007		
Metric	331.735**	42	.993	.013	-.006	.006
Scalar	2520.93**	74	.941	.029	-.052	.016
Extreme Response Style Corrected						
Configural	53.526**	18	.999	.007		
Metric	325.068**	42	.993	.013	-.006	.006
Scalar	486.821**	74	.990	.012	-.003	-.001
Overclaiming Corrected						
Configural	53.611**	18	.999	.007		
Metric	328.073**	42	.993	.013	-.006	.006
Scalar	636.253**	74	.987	.014	-.006	.001
Anchored Scores						
Configural	91.073**	18	.999	.010		
Metric	267.398**	42	.997	.012	-.002	.002
Scalar	2601.131**	74	.972	.030	-.025	.018

** $p < .01$.

Although the anchored scores did not achieve scalar invariance, the model fit of the anchored scores was better than that of the raw scores, indicated by the change of CFI value from metric to scalar invariance model of .25 and .52 in these two cases respectively. To summarize, controlling for extreme response style or overclaiming increased the comparability of scores in these nine countries. Anchoring vignettes improved the comparability to some extent, but did not yield full comparability.

Mean Patterns Before and After Correction

The latent mean scores of the teacher support scale for each country was estimated in the multigroup confirmatory factor analysis in all four cases. Mexico was treated as the reference group, because of the largest sample size in this country. Technically, the latent mean of Mexico was constrained to be zero in the scalar invariance model, and the latent means of other countries were freely estimated. The comparison of the mean patterns with the 95% confidence intervals has been plotted in Figure 1. Note that the scores were not reverse-coded; therefore the means in this Figure represented level of lack of teacher support.

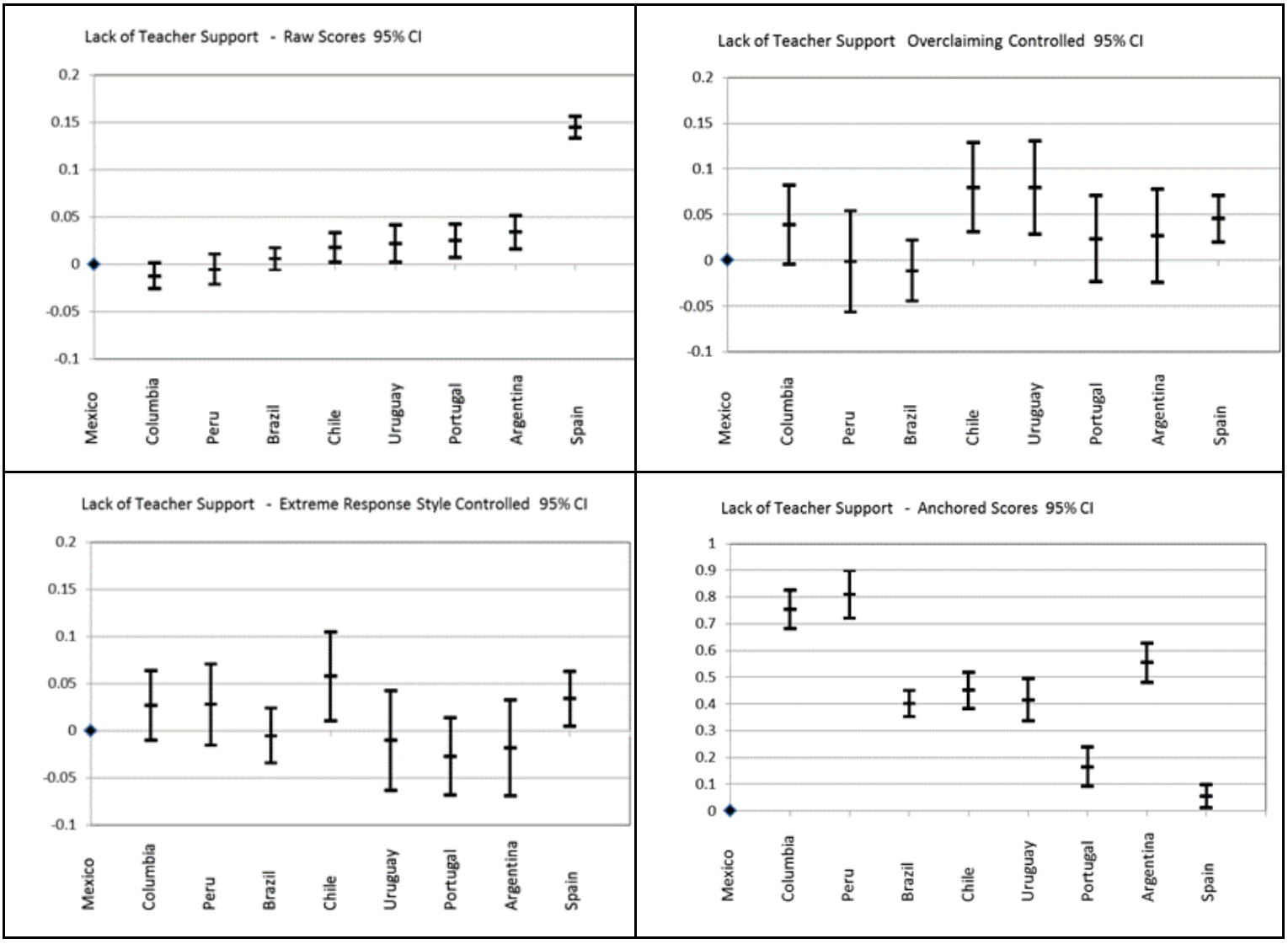

Figure 1 *Mean patterns in raw and adjusted scores on lack of teacher support (Mexico as reference country)*

With raw scores, most of the Latin American countries did not differ much on this construct, except that Spain showed a lower level of teacher support compared with all other countries. When extreme response style was corrected for, the mean pattern changed in two noticeable ways. Firstly, the confidence intervals all increased, indicating that more measurement errors had to be taken into consideration. Secondly, the difference between Spain and other countries became much smaller due to the correction. A similar pattern was observed when overclaiming was controlled for. In all these above mentioned cases, country differences in teacher support were rather limited, whereas the pattern was drastically changed when the anchored scores were used. With anchored scores, Columbia and Peru showed a significantly lower level of teacher support, compared with Spain, Portugal, and Mexico. In other words, the correction effects on mean patterns were rather different, with the anchoring vignette approach showing the largest change and the other two approaches more limited change.

Correlations Before and After Correction

It was expected that teacher support would correlate positively with teacher-directed instruction. The expectation about the correlation between teacher support and math achievement is less clear. On the one hand, these scales should be positively correlated, given that positive interactions with teachers contribute to better learning. On the other hand, students who perceived most teacher support might be the ones who did not perform well, thus a negative correlation is not unreasonable. The zero-order correlations with raw factor scores and anchored factor scores of teacher support, and partial correlations with extreme response style and overclaiming controlled for are presented in Table 3.

Table 3 *Correlations of Teacher Support in Each Country*

	Correlation with Teacher-Directed Instruction				Correlation with Math Achievement (PV1)			
	Raw	ERS Adjusted	Overclaim Adjusted	Anchored	Raw	ERS Adjusted	Overclaim Adjusted	Anchored
Argentina	.599	.561	.601	.121	-.044	*-.041*	-.073	.134
Brazil	.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
Spain	.648	.642	.653	.264	-.027	*-.029*	-.028	.136
Mexico	.612	.567	.608	.150	-.028	-.056	-.045	.146
Peru	.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

Note. The correlations with math achievement were carried with the five plausible values, and all showed the same results, thus only correlations with the first plausible value (PV1) were reported. ERS= extreme response style. All correlations are significant at $p < .01$ except the italicized ones.

In all four cases, teacher support and teacher-directed instruction were positively related. The correction based on extreme response style and overclaiming had a rather limited effect; the slight reduction in correlations suggested that some general scale usage preference was partialled out. The change in size of correlation was more salient in anchored scores. The correlations with math achievement when the raw score, extreme response style correction, and overclaiming correction that were used were slightly negative in general, whereas with anchored scores, teacher support was positively related to math achievement. It pointed to the effectiveness of anchoring vignettes in reversing the correlations between positive experience in learning and achievement.

However, whether the anchored scores are more valid is still not clear.

Discussion

We studied the effects of three correction methods on self-report Likert scales in nine countries with similar scale usage preferences with the 2012 PISA data. These target countries (Latin American and South European countries) generally show a strong expressiveness and self-enhancement tendency, which might influence the validity of self-reports. We examined the impact of corrections for extreme response style, overclaiming, and anchoring vignettes on measurement invariance, country mean, and correlation with external variables. The main findings include: (1) All corrections helped improve the measurement invariance level, although anchored scores were less effective in reaching scalar invariance compared with the correction of extreme response style and overclaiming; (2) controlling for extreme response style and overclaiming had limited effects on country mean scores or correlation with other variables, whereas anchored scores showed more drastic changes. There is no evidence showing which correction actually works best in enhancing the validity of scores in these countries. The improvement in invariance statistics using notably the extreme response style and overclaiming corrections suggest that score corrections may enhance the validity. In addition, the plot of the country means after corrections for these biases has more intuitive appeal in that the Spanish country mean is now closer to the mean of the other countries; this pattern seems more plausible as we are not aware of literature showing that Spanish teachers support their students much less than teachers in the other countries. However, the lack of any impact of these corrections on the correlations is counterintuitive. A large score on the global extreme response style measure should also be present in the teacher support score (correlation of these two is .25), which should lead to a considerable reduction of the score. These correlations are more affected by anchoring. However, there is no evidence that these correlations are more realistic (neither for teacher-directed instruction nor for Math achievement).

The potential validity threats from differential scale usage preferences in Latin American countries loom large, especially in large-scale assessment contexts, where cross-cultural comparative data are used to inform evidence-based policy making (e.g., Goldstein, 2004; Gorur, 2014). In PISA, the paradoxical reversal of individual- and country-level correlations between self-report positive experiences and achievement created the necessity to correct for scale usage preferences. However, in the present study, the correction with different methods in seven Latin American and two Southern European countries showed rather mixed results: correction for extreme response style and overclaiming ensured full scalar invariance in the teacher support scale, but did not change correlation patterns with external variables; Whereas anchoring vignettes changed correlation patterns but did not help reach scalar invariance. It seems that these correction methods target different scale usage preferences. Extreme response style and overclaiming are more general scale usage preferences that affect all kinds of self-reports in a uniform way. Anchoring vignettes target the individual differences in interpreting the content and response options in more specific ways. The rescaling based on anchoring vignettes may bring out more variation in scores at both individual and cultural level. However, it is not clear whether the rescaling introduces other type of bias, in particular, given the likely violation of the (stringent) assumptions of anchoring vignettes. It can be concluded that score corrections such as derived from anchoring vignettes are capable to reverse the motivation—achievement paradox when comparing widely different regions, such as East Asia and Latin America, these procedures yield results that are much more difficult to interpret when applied in a culturally more homogenous region.

There are many procedures that presumably enhance the comparability and validity of cross-cultural self-report data. It is difficult to tease stylistic responding out from the substantive construct being measured, because scale usage preferences might be an integral part of respondents' psychological makeup (He & van de Vijver, 2015c), and aggregated at culture level, they may represent important aspects of national culture, such as individualism—collectivism value preferences (Smith, 2004, 2011). Some studies demonstrate significant correction effects on country comparisons (e.g., Diamantopoulos, Raeynolds, & Simintiras, 2006), whereas some other studies report negligible effects (e.g., He & van de Vijver, 2015a). The findings of our study show inconsistent findings across correction methods. The question of which correction method indeed reduces bias and enhances validity on different measures decisively awaits further research efforts.

In a practical sense, it is still worth the effort to compare scores with and without various corrections. As both extreme response style and overclaiming function similarly in terms of their correction effects, and with the consideration that extreme response style can be constructed with various existing item responses whereas overclaiming requires an additional measure, it seems that correction for extreme response style is easier to implement. The use of anchoring vignettes requires caution; no conclusive findings can be reported with anchored scores until the two assumptions of this method are empirically tested and satisfied.

Limitations and Future Directions

This study has a few limitations. Firstly, we only targeted nine countries out of the 64 PISA countries. This is because the self-enhancement and expressiveness tendency in these target countries are particularly worrisome in overestimation of self-reports. Further studies can broaden the search to more varied cultural contexts. Secondly, we restricted our analysis to students who answered the Form C of the student questionnaire, in order to avoid a bulk amount of missing values. Therefore the country means estimated are based on approximately one third of the total sample, which may not be entirely nationally representative. Lastly, we limited our correction methods given the availability of data. There are other item design methods such as forced-choice format questions and situational judgement format questions and other statistical corrections such as bi-factor models that may help remedy the lack of comparability and validity concerns (e.g., Brown & Maydeu-Olivares, 2011; Cheung & Rensvold, 2000; Rutkowski et al., 2014). As more efforts are put in alleviating measurement bias in different cultural contexts, we believe large-scale assessment data can be better utilized for basic research and evidence-based policy making.

References

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. doi: 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*. 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/S000305540400108X

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

Notes

[1] Only in Form C of the student background questionnaire have the target measures (teacher support, overclaiming, and anchoring vignettes on teacher support) been administered. This sub-sample allows less biased sample means be estimated.

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

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

Ph.D Cross-Cultural Psychology, Tilburg University, Tilburg, The Netherlands, 2011-2015. Master of Arts, Intercultural Communication, Shanghai International Studies University, Shanghai, China. Bachelor of Arts, Marketing, Dongbei University of Finance and Economics, Dalian, China. Post Doc Deutsches Institut für Internationale Pädagogische Forschung (DIPF) since 2015. Research on self-report data comparability in large-scale international surveys, using advanced psychometric methods and new item formats, etc. Research World Bank Group Consultant since 2015. Analyst OECD since 2015. Postal Address: Department of Educational Quality and Evaluation, Schloßstraße 29. 60486 Frankfurt am Main (Germany).

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

Professor of Cross-Cultural Psychology at the Tilburg University, the North-West University and the University of Queensland, known for his work on cross-cultural research and on "methods and data analysis of comparative research". Van de Vijver received both his MA and in 1991 his PhD in Psychology at the Tilburg University. He is appointed Professor cross-cultural psychology at Tilburg University, and is also Professor at the North-West University in South Africa and the University of Queensland in Australia. In 2013 he and Maria Cristina Richaud received the APA Award for Distinguished Contributions to the International Advancement of Psychology.

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

[ISSN: 1134-4032]