National Academy of Education

Workshop Series on Methods and Policy Uses of International Large-Scale Assessments (ILSA)

A Look at the Most Pressing Design Issues in International Large-Scale Assessments

Leslie Rutkowski
Centre for Educational Measurement
University of Oslo, Norway

December 2016

Contact: Leslie Rutkowski, Centre for Educational Measurement at University of Oslo, Postboks 1161 Blindern, 0318 OSLO Norway, +47 22 84 44 90, leslie.rutkowski@cemo.uio.no.

This paper was prepared for the National Academy of Education's *Workshop Series on Methods and Policy Uses of International Large-Scale Assessment (ILSA)*. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305U150003 to the National Academy of Education. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

A LOOK AT THE MOST PRESSING DESIGN ISSUES IN INTERNATIONAL LARGE-SCALE ASSESSMENT

Abstract

Three pressing design issues in international large-scale assessments, such as Progress in International Reading Literacy, Programme for International Student Assessment, and Trends in International Mathematics and Science Study, are outlined. In all three cases, the importance of the matter at hand and proposed solutions are set against the backdrop of educational policy. The first matter regards issues around cultural comparability of the test and context questionnaires. Cultural modifications to current studies are recommended. The second topic takes up the presence and problem of measurement error, particularly in context questionnaires, and the way that less error-prone measures might be collected in the case of key reporting variables. The final topic deals with the desire to draw causal inferences from international assessment data and the challenges therein under current designs. Two proposals are offered for strengthening the foundation upon which select causal effects might be estimated. Although none of the proposed solutions are trivial, each one offers the possibility of better meeting the demands placed on international assessments in a modern, globalized, and highly heterogeneous world.

INTRODUCTION

In the past 20 or so years, international assessments have come of age and assumed a prominent place in educational policy and research discussions. Particularly in the case of the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Assessment (PISA), results have stimulated considerable changes in a number of participating educational systems.[1] Perhaps the best-known among these are the massive, swift educational reforms in Germany that were stimulated by the 2000 PISA results and ensuing "PISA-shock" (Ertl, 2006). Similar effects were felt in a number of European countries including Denmark (Egelund, 2008), Finland (Dobbins & Martens, 2012), and others (Grek, 2009). In the United States, PISA results have been likened to *Sputnik* (Finn, 2010); they have also formed the basis for calls to improve U.S. educational standards (Duncan, 2013). And although this commissioned paper emphasizes the most pressing methodological issues related to design in international assessment, policy and research issues are inextricably linked to the models and methods of international assessment. As such, I take up a discussion of pressing design issues in international large-scale assessments (ILSAs) against a backdrop of educational policy.

Efforts of the organizations responsible for the prominent international assessments—the IEA and the OECD—have been monumental. They and their contractors have contributed in meaningful ways to an international understanding of achievement patterns. The work of these organizations has also facilitated educational measurement

[1] Given the complexity and ambiguity in conceptions of the "nation-state," particularly for city-states, non-national systems, or territories with disputed or ambiguous political status, we refer to PISA participating units as *educational systems*. Examples of geographic areas with special status include Dubai (an emirate within the United Arab Emirates), Taiwan (a geographic area with a disputed or ambiguous political status), Hong Kong and Macao (special administrative regions of China), and Singapore (a city-state).

and comparison in systems without the internal capacity for such studies. And the state of the art continues to move forward with each cycle, as innovations are piloted and adopted. Nevertheless, every endeavor can be improved. In what follows, several areas that are relevant to the overarching topic are outlined. This discussion is supported with examples from my own research as well as the work of other psychometricians and methodologists working in the field of international assessment. Although this paper is primarily a review and synthesis, it also contains new evidence from recent cycles of international assessments to highlight areas of importance. The paper begins with problems around cross-cultural comparability of achievement and non-achievement measures and the way in which the field might address some of the underlying issues. Statistical and design-based solutions are proposed as possible ways of moving forward. This is followed by a discussion around data quality and measurement error, particularly as they pertain to reporting achievement differences on key policy-relevant variables. Several possible ways that assessment design could be brought to bear on the problem are suggested. The paper concludes by describing a final issue that links ILSA design with data use: causal inferences and the observational, cross-sectional nature of international assessments. One study that would serve as a testbed for developing solutions to this final issue is indicated. Finally, given the mandate of this invited paper, model-based problems or solutions are not taken up, but rather the discussion is limited to *design issues* and associated solutions. This omits, for example, discussions around recent methodological innovations in partial invariance spearheaded by the Mplus group (Asparouhov & Muthén, 2014).

CROSS-CULTURAL COMPARABILITY

ILSAs such as the Trends in International Mathematics and Science Study (TIMSS) and PISA are tasked with measuring what students know and can do internationally. Furthermore, non-performance-based education studies, such as the Teaching and Learning International Survey (TALIS), seek to measure study participants on latent variables that deal with attitudes, perceptions, and experiences. In both types of study, performance on tasks or scores on other latent variables are typically summarized in terms of measurement model-based scale scores (Martin & Mullis, 2012; OECD, 2014b). Regardless of the survey topic or specific data source, an important criterion for comparing scale scores in an international context is that the latent variable is understood and measured equivalently across all countries. This property is typically (although not *only*) referred to as *measurement invariance* or *differential item functioning* (DIF). Generally, investigations of measurement invariance focus on the degree to which comparisons on the latent variable of interest (e.g., teacher beliefs) can be validly compared across heterogeneous populations. To ensure that international and national assessments meet their goals for all stakeholders, it is critical that country- and subpopulation-level achievement (e.g., boys compared to girls, native-born versus immigrant students) is correctly and precisely estimated. Furthermore, achievement ranking schemes must be appropriate, while also recognizing that rankings are imperfect.

Although equivalence-of-scale scores in large-scale achievement tests have received substantial attention in the academic literature (e.g., Allalouf, Hambleton, & Sireci, 1999; Ercikan, 2002, 2003; Grisay & Monseur, 2007; Hambleton, Merenda, & Spielberger,

2005), a relatively recent extension of scale score equivalence to non-achievement-scale scores has emerged (OECD, 2014b,c). To that end, the 2008 and 2013 cycles of TALIS and the 2012 cycle of PISA used multiple-groups confirmatory factor analysis (MG-CFA; Jöreskog, 1971) to provide evidence of score comparability on a number of scales designed to measure and compare students and teachers internationally in areas such as beliefs and practices (OECD, 2010). OECD researchers supported their findings regarding measurement invariance with research that was limited in scope to few groups and relatively small sample sizes (Chen, 2007; French & Finch, 2006). In general, findings suggested that strong invariance of the sort necessary to compare means was rarely achieved, obviating the possibility of direct comparisons across countries.

Despite the importance of optimal, unbiased estimates of latent traits such as achievement or affective domains like motivation, ILSAs have been found to suffer from limitations in this regard. In particular, the statistical methods used to estimate system-level achievement have been shown to suffer from some deficiencies (Brown et al., 2007; Goldstein, 2004; Mazzeo & von Davier, 2009; Oliveri & von Davier, 2011, 2014). I review recent studies that directly investigate this issue and that propose reasonable solutions. Added to this discussion is a proposal to add *culturally* or *regionally* specific items or clusters of items to enhance cross-cultural comparability of international assessments. First, however, the models used to estimate achievement in international assessments such as PISA, TIMSS, and PIRLS are briefly described.

Background on Achievement Estimation

ILSA programs employ sophisticated test booklet designs whereby each individual student is administered just a small number of the total possible items, yet all items are administered throughout each of the reporting groups. As one example, more than 10 hours of testable material was available for the TIMSS 2011 assessment (Mullis et al., 2009). To minimize individual examinee burden, test developers used an assessment design that distributed the total test content into 14 non-overlapping mathematics blocks and 14 non-overlapping science blocks. That is, the blocks exhaustively and mutually exclusively contained all available testing material. These blocks subsequently were arranged into 14 booklets containing 2 science and 2 mathematics blocks each, with no blockwise overlap within a booklet. As such, no block would appear more than once within a booklet. This design ensured linking across booklets because each block (and therefore each item) appeared in two different booklets. Furthermore, the total assessment material was divided into more reasonable 90-minute periods of testing time for each student.

This approach to item administration is often referred to as item sampling (Lord, 1962) or, more commonly in current ILSA literature, as multiple-matrix sampling (Shoemaker, 1973). To overcome the methodological challenges associated with multiple-matrix sampling (Mislevy, Johnson, & Muraki, 1992), ILSA programs adopted a population or latent regression modeling approach that uses marginal estimation techniques to generate population- and subpopulation-level achievement estimates (Mislevy, 1991; Mislevy et al., 1992). Under the latent regression modeling approach, consistent population- and subpopulation-level ability estimates are achieved by treating achievement as missing (latent) data. These data points are missing for all examinees and are "filled

in" using an approach analogous to multiple imputation (Rubin, 1976, 1987). As in multiple-imputation methods, an imputation model (called a "conditioning model") is developed to estimate posterior population achievement distributions. This model uses all available student data (cognitive as well as background information) to generate a conditional proficiency distribution for each student from which to draw plausible values (usually five) for each student on each latent trait (e.g., mathematics, science, and associated subdomains). Subpopulation estimates of achievement derived from the conditioning models used in this approach are less biased than those estimated via traditional item response theory (IRT) methods (Mislevy, 1991; Mislevy et al., 1992; von Davier, Gonzalez, & Mislevy, 2009).

Because achievement (θ) is a latent, unobserved variable for every examinee, it is reasonable to treat θ as a missing value and to approximate statistics involving θ by its expectation. That is, for any statistic t, $\hat{t}(\mathbf{X},\mathbf{Y}) = \mathrm{E}[t(\theta,\mathbf{Y})[\mathbf{X},\mathbf{Y}] = \int t(\theta,\mathbf{Y})p(\theta \mid \mathbf{X},\mathbf{Y})d\theta$, where $\mathbf{X}$ is a matrix of item responses for all examinees and $\mathbf{Y}$ is the matrix of responses of all examinees to the set of administered background questions. Because closed-form solutions are typically not available, random draws from the conditional distributions $p(\theta \mid \mathbf{x}_i,\mathbf{y}_i)$ are drawn for each sampled examinee, i (Mislevy, Johnson, & Muraki, 1992). These are typically referred to as plausible values in ILSA terminology or *multiple imputations* in missing data literature. Using Bayes' theorem and the IRT assumption of conditional independence,

$$\begin{aligned} p(\theta \mid \mathbf{x}_i,\mathbf{y}_i) &\propto P(\mathbf{x}_i \mid \theta,\mathbf{y}_i)p(\theta \mid \mathbf{y}_i) \\ &= P(\mathbf{x}_i \mid \theta)p(\theta \mid \mathbf{y}_i), \end{aligned} \tag{1}$$

where $P(\mathbf{x}_i \mid \theta)$ is the likelihood function for θ induced by observing $\mathbf{x}_i$, and $p(\theta \mid \mathbf{y}_i)$ is the distribution of θ for a given vector of response variable. The distribution of θ is assumed normal with a mean given by the following linear model (the *conditioning* model) such that $\mathbf{y}^c$ is the vector of (usually assumed) *perfectly measured* background variables,

$$\theta = \mathbf{\Gamma}'\mathbf{y}^c + \in \tag{2}$$

where $\in \sim N(\mathbf{0},\mathbf{\Sigma})$ and $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ are estimated. Operationally, student background variables and some system-specific variables are subjected to a principal component analysis. The resulting principal components are used as predictors in the conditioning model. This has the effect that several hundred background variables are reduced to several dozen predictors that are linear combinations of the original variables.

Under an IRT framework, meaningful cross-cultural comparisons depend on item parameter equivalence (Hambleton & Rogers, 1989; Mellenbergh, 1982; Meredith, 1993; Millsap, 2011). This implies that test items are assumed to be equally difficult across the populations under consideration. That is, an item should be equally difficult for students in Kazakhstan, Norway, and Shanghai. Because operational procedures in ILSAs rely on this assumption and detailed analyses are conducted by the relevant organizations to identify and ameliorate issues with *item-by-country* interactions (Martin & Mullis, 2012; OECD, 2014b), it is notable that in empirical investigations, the assumption does not hold (e.g., Kreiner & Christensen, 2014; Oliveri & von Davier, 2011; Rutkowski, Rutkowski, & Zhou, 2016). And in a limited investigation (Rutkowski et al., 2016), violations were

found to have consequences especially for ranking middle-performing educational systems. Furthermore, the same study showed that achievement can be meaningfully biased—in several cases resulting in achievement outside of the original 95 or 99 percent confidence interval, pointing to the importance of measurement equivalence in an ILSA setting.

Recent Investigations

Three recent studies investigated the potential for allowing cross-cultural differences in the models used to estimate scale scores for achievement (von Davier, 2015) and non-achievement scales, the items of which were administered in PISA 2009 and 2012, respectively (Glas & Jehangir, 2014; Rutkowski & Rutkowski, 2016). Across all three studies, the findings suggested that permitting some cultural specificity in the models used to estimate scale scores produced better model-data consistency. In particular, relaxing the strict assumption of parameter equivalence showed promising results, as did allowing for some country-specific items that better tapped into the local context. The latter point is further developed as a possible addition to future rounds of international assessments.

Tailoring Items

An innovation in the 2009 and 2012 cycles of PISA was the option of including easier booklets into the assessment for educational systems with low expected performance. Countries that chose not to include easier booklets were administered the "standard" test. This effort, incorporated only into the math portion of the 2012 assessment, was intended to better capture what students in low-performing countries know and can do (OECD, 2014b, p. 31). The PISA 2012 booklet design is represented in Table 1. Content clusters are denoted by a letter and a number, where M, S, and R indicate math, science, and reading, respectively. Regardless of a country's choice to participate in the standard booklet or easier booklet administration, 13 booklets were administered in a rotated, random fashion so that each booklet would be administered an approximately equal number of times within a country. Note that of nine total math clusters, M6A and M7A were the optional standard clusters, while M6B and M7B were the optional easier clusters. No other math clusters were subject to this choice. This design resulted in booklets 1 to 7 corresponding to the standard option while booklets 21 to 27 corresponded to the easy option. Furthermore, booklets 8 to 13 represented the core booklets and were administered in all countries to ensure sufficient test material for linking.

TABLE 1. PISA 2012 Booklet Design

Designation	Booklet ID	Cluster				Standard Booklet Set	Easier Booklet Set
Standard only	1	M5	S3	**M6A**	S2	X	
	2	S3	R3	**M7A**	R2	X	
	3	R3	**M6A**	S1	M3	X	
	4	**M6A**	**M7A**	R1	M4	X	
	5	**M7A**	S1	M1	M5	X	
	6	M1	M2	R2	**M6A**	X	
	7	M2	S2	M3	**M7A**	X	
Core	8	S2	R2	M4	S1	X	X
	9	R2	M3	M5	R1	X	X
	10	M3	M4	S3	M1	X	X
	11	M4	M5	R3	M2	X	X
	12	S1	R1	M2	S3	X	X
	13	R1	M1	S2	R3	X	X
Easy only	21	M5	S3	**M6B**	S2		X
	22	S3	R3	**M7B**	R2		X
	23	R3	**M6B**	S1	M3		X
	24	**M6B**	**M7B**	R1	M4		X
	25	**M7B**	S1	M1	M5		X
	26	M1	M2	R2	**M6B**		X
	27	M2	S2	M3	**M7B**		X

NOTE: Clusters marked in bold are those that are optionally easier or standard.

Although OECD can be commended for recognizing that low-performing educational systems are not well measured by the standard PISA design, it can also be argued that ceiling effects are similarly important and are probable in this context. As an example, a review of the PISA 2012 *Compendium for the Cognitive Item Response*s (OECD, n.d.) revealed that in high-performing systems such as Finland, Shanghai, and Singapore a far greater proportion of students correctly answered many of the mathematics items than in the total international sample. In some cases, these differences were stark (e.g., 0.62 internationally and 0.88 in Singapore on item PM915Q02; 0.56 internationally compared to 0.71 in Finland on item PM205Q01; and 0.78 internationally versus 0.94 in Shanghai on item PM423Q01). As a further example, I examined the proportion of math items that were correctly answered by at least 80 percent of examinees. In Shanghai and Finland that figure was 0.350 and 0.138, respectively. In contrast, the proportion of easy items (by an arbitrary cutoff of 0.80) was 0.057 internationally and 0.025 in Chile (an OECD country that opted for the easy booklet option). Considered from another perspective, the proportion of items where 20 percent or fewer answered correctly was also examined. In Shanghai and Finland the proportion was 0.025 and 0.125, respectively. Internationally and in Chile, the proportions were 0.125 and 0.288, respectively. Notably for Chile, these findings include items in the easy booklets.

The previous examples are not comprehensive; however, they are representative of a trend among the highest-performing systems. That is, at the item level, these systems are vastly outperforming the full international sample. Similar results can be seen when proficiency levels are taken into account. PISA defines six proficiency levels, where 1 is the lowest level of performance and 6 is the highest level (defined as 669 points or higher on the PISA math scale OECD, 2014a). In Shanghai, 30.8 percent of students are at or above level 6. Interestingly in this case, no other proficiency level has a higher percentage of Shanghai students. Several other participating systems exhibited high percentages of students performing at level 6, including Singapore (19.0 percent), Taiwan (18.0 percent), and Hong Kong (12.3 percent). In contrast, the OECD average percentage of level 6 performers was just 3.3 percent, again pointing to evidence that the very highest achieving systems might be better measured by including more difficult item clusters into the assessment design.

In light of these simple comparisons, it should be relatively straightforward, then, to argue in favor of a *hard/challenging* booklet option for top performers to ensure that the instrument captures the full spectrum of proficiency in these educational systems. These booklets could be designed and administered in the same way as the easy booklets, featuring two item clusters with more difficult items and a booklet design that mimics the standard-only/easy-only distinction illustrated in Table 1. Although this would pose additional operational challenges for item calibration and linking scores across the booklets, the methods that are already well established for the easy booklets could readily be extended for the challenging booklet option.

Both design- and model-based solutions could also be applied to alleviate some of the concerns around background questionnaires. The first, and perhaps most obvious, is for countries to make better use of the national option that allows for the inclusion of country-specific questions, extending the design-based solutions to background scales. Unfortunately, not all countries that participate in ILSAs include these national options and when they do, country-specific items are normally used sparingly. Because national options can provide insights into educational systems that are not available in the more general international questionnaire, the capacity to explain differences in achievement can be significantly increased by introducing more national options. That said, developing valid and reliable background questionnaire items is a task that should not be taken lightly and can require significant resources. Regions who share many commonalities could work together to help develop questions and scales targeted toward their systems. Furthermore, the test developers could adopt more flexible methods for incorporating these unique items into the background scales to ensure that each participating system is well measured on the construct of interest, while maintaining comparability across systems.

IMPROVING THE MEASUREMENT OF KEY REPORTING VARIABLES

One important function of ILSAs is to report overall achievement within and across systems. Another equally important purpose is to disaggregate achievement across policy-relevant subgroups of examinees. For example, both PISA and TIMSS report average achievement for boys and girls in all content domains. In addition, both studies report achievement differences across levels of socioeconomic status (SES) and other variables associated with student background (Mullis et al., 2012; OECD, 2014a). Importantly, however, such measures, collected from the students and/or parents are self-reported and therefore particularly susceptible to different forms of measurement error (He & van de Vijver, 2013). Furthermore, there is evidence to suggest that measurement error is worse in less-economically developed educational systems (Rutkowski & Rutkowski, 2010, 2016), where both studies examined response consistency between fourth-grade students and their parents. This perspective is substantiated by Hauser's (2013) review of the PISA SES measure, where it is noted that meaningful cross-national differences in reliability exist, leading to variable impacts on the attenuation of relationships involving this measure. Finally, there is meaningful bias in subgroup achievement estimates, the magnitude of which depends directly on the type and magnitude of measurement error in these variables (Rutkowski, 2014).

Subsequently, I consider an example from an older population of examinees and their parents. In PISA 2012, educational systems were offered the opportunity to administer a parent questionnaire along with the usually administered student and school master questionnaires. And between the two questionnaires, there is one common question regarding whether the child in question had ever repeated a grade at ISCED 1 (elementary school), 2 (middle school), or 3 (high school). In both cases, response options include "*no, never*"; "*yes, once*"; and "*yes, twice or more*." Of the PISA participating systems, 11 countries administered these questions to both parents and children. The results of countrywise cross-classification of these two variables can be found in Table 2, which is populated with unweighted proportions. In general, as expected, students and their parents respond with a high degree of consistency on these two items. That is, most observations fall along the diagonal of the tables for each country (marked in gray). Nevertheless, in nearly every participating country and at all three ISCED levels, a pattern is present, such that either parents respond with "*never*" and their children omit a response, or vice versa (marked in rose). Notably, at ISCED 3 in Hong Kong and Portugal, nearly 20 percent of data are a *never/omit* combination. This suggests that 20 percent of students and their parents respond differently, with one or the other electing to omit a response. Again, the reasons why are, given the available information, not known.

TABLE 2. Cross-Classification of Parent and Student Responses on the PISA 2012 Question "Have you/Has your child ever repeated a grade"?

Country	Parent Response	ISCED 3 Student Response				ISCED 2 Student Response				ISCED 1 Student Response			
		Never	Once	More	Miss	Never	Once	More	Miss	Never	Once	More	Miss
Belgium	Never	.82	.00	.00	.06	.85	.01	.00	.05	.81	.01	.00	.03
	Once	.00	.03	.00	.00	.00	.04	.00	.00	.01	.10	.00	.01
	More	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	Miss	.05	.00	.00	.03	.02	.00	.00	.01	.01	.00	.00	.01
Chile	Never	.75	.01	.00	.06	.80	.00	.00	.07	.80	.00	.00	.04
	Once	.00	.06	.00	.00	.00	.03	.00	.00	.00	.06	.00	.01
	More	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00
	Miss	.05	.00	.00	.06	.02	.00	.00	.06	.02	.00	.00	.05
Croatia	Never	.87	.01	.00	.03	.96	.01	.00	.02	.96	.01	.00	.01
	Once	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	More	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	Miss	.06	.00	.00	.01	.01	.00	.00	.00	.01	.00	.00	.00
Germany	Never	No valid responses				.80	.01	.00	.05	.81	.01	.00	.05
	Once					.01	.08	.00	.00	.01	.06	.00	.01
	More					.00	.00	.00	.00	.00	.00	.00	.00
	Miss					.02	.00	.00	.02	.02	.00	.00	.03
Hong Kong	Never	.76	.00	.00	.09	.86	.00	.00	.04	.85	.02	.00	.03
	Once	.00	.00	.00	.00	.00	.06	.00	.00	.01	.06	.00	.00
	More	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	Miss	.09	.00	.00	.05	.02	.00	.00	.00	.02	.00	.00	.01

Hungary	Never	.84	.01	.00	.03	.92	.01	.00	.02	.92	.01	.00	.02
	Once	.00	.01	.00	.00	.00	.01	.00	.00	.00	.01	.00	.00
	More	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	Miss	.07	.00	.00	.03	.02	.00	.00	.00	.02	.00	.00	.01
Italy	Never	.84	.00	.00	.02	.86	.00	.00	.03	.88	.00	.00	.04
	Once	.00	.07	.00	.00	.00	.04	.00	.00	.00	.00	.00	.00
	More	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00
	Miss	.02	.00	.00	.02	.02	.00	.00	.02	.03	.00	.00	.04
Korea	Never	.85	.01	.00	.02	.94	.02	.01	.01	.94	.02	.01	.00
	Once	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	More	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	Miss	.08	.00	.00	.02	.01	.00	.00	.00	.01	.00	.00	.00
Macedonia	Never	.55	.00	.00	.12	.60	.01	.00	.04	.64	.02	.00	.04
	Once	.01	.00	.00	.00	.01	.21	.01	.01	.01	.12	.01	.00
	More	.00	.00	.00	.00	.00	.00	.03	.00	.00	.00	.04	.00
	Miss	.15	.00	.00	.17	.04	.01	.00	.03	.05	.01	.00	.04
Mexico	Never	.71	.00	.00	.08	.87	.00	.00	.06	.86	.01	.00	.02
	Once	.01	.00	.00	.00	.01	.01	.00	.00	.01	.06	.00	.00
	More	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00
	Miss	.05	.00	.00	.14	.02	.00	.00	.02	.01	.00	.00	.01
Portugal	Never	.53	.00	.00	.16	.09	.00	.00	.01	.09	.00	.00	.01
	Once	.00	.00	.00	.02	.00	.01	.00	.00	.00	.02	.00	.00
	More	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	Miss	.04	.00	.00	.24	.00	.00	.00	.01	.00	.00	.00	.01

As a final example, results of a simple analysis based on PIRLS 2011 are included, which featured both a parent and a student questionnaire in all participating educational systems (Mullis et al., 2009). There are a few questions that are the same or similar between both groups of respondents, and Table 3 includes the correlation (raw and unattenuated) between responses on a question about the number of books in the home (a historical proxy for SES and generally predictive of achievement internationally). Although correlations are generally high between parent and student responses, particularly after the correlations are corrected for measurement error, there are still several relatively low correlations that tend to be concentrated in economically developing countries (excepting Malta, which has a low correlation but a highly industrialized economy). This suggests a meaningful level of disagreement in many countries and educational systems.

TABLE 3. Correlation Between Parent and Student Report of Number of Books in the Home

	Raw Correlation		Unattenuated Correlation			Raw Correlation		Unattenuated Correlation	
Country	r	SE	r_{corr}	SE_{corr}	Country	r	SE	r_{corr}	SE_{corr}
Indonesia	0.25	0.03	0.26	0.01	Dubai	0.48	0.01	0.79	0.02
Kuwait	0.25	0.03	0.35	0.03	Italy	0.49	0.02	0.73	0.03
Republic of Azerbaijan	0.28	0.02	0.35	0.02	Iran	0.49	0.02	0.69	0.02
Qatar	0.29	0.02	0.46	0.03	Poland	0.49	0.01	0.64	0.02
Malta	0.30	0.02	0.42	0.03	Slovenia	0.49	0.01	0.60	0.02
Morocco	0.31	0.02	0.34	0.02	Hong Kong	0.49	0.02	0.68	0.03
Botswana	0.33	0.03	0.36	0.02	France	0.50	0.02	0.73	0.03
Abu Dhabi	0.36	0.02	0.52	0.03	Belgium (French)	0.50	0.02	0.78	0.03
Trinidad and Tobago	0.38	0.02	0.53	0.03	Finland	0.50	0.02	0.62	0.02
Canada (Alberta)	0.38	0.02	0.52	0.03	Morocco (Grade 6)	0.51	0.02	0.51	0.01
United Arab Emirates	0.40	0.01	0.66	0.02	Spain (Andalucia)	0.51	0.01	0.69	0.02
Oman	0.40	0.01	0.56	0.02	Czech Republic	0.51	0.02	0.68	0.02
Singapore	0.43	0.01	0.57	0.02	Germany	0.51	0.02	0.74	0.03
Canada	0.44	0.01	0.53	0.01	Ireland	0.51	0.02	0.80	0.03
Canada (Quebec)	0.44	0.02	0.57	0.02	Spain	0.51	0.02	0.71	0.02
Norway	0.44	0.02	0.62	0.03	Georgia	0.55	0.02	0.92	0.03
Australia	0.45	0.02	0.63	0.03	Austria	0.55	0.01	0.82	0.03
Canada (Ontario)	0.45	0.02	0.57	0.02	Lithuania	0.55	0.01	0.71	0.02
Saudi Arabia	0.45	0.02	0.78	0.03	South Africa	0.55	0.03	0.75	0.03
Republic of Honduras	0.45	0.03	0.49	0.02	Portugal	0.56	0.02	0.76	0.03
Netherlands	0.46	0.02	0.65	0.03	Chinese Taipei	0.56	0.01	0.91	0.03
Int. Avg.	0.46		0.66		Sweden	0.57	0.02	0.86	0.03
Northern Ireland	0.47	0.02	0.72	0.03	Croatia	0.57	0.01	0.70	0.02
Israel	0.47	0.02	0.71	0.03	Denmark	0.59	0.01	0.84	0.03
New Zealand	0.47	0.02	0.70	0.03	Romania	0.62	0.02	1.00	0.03
Colombia	0.48	0.02	0.48	0.02	Slovak Republic	0.65	0.02	0.90	0.02
Russian Federation	0.48	0.02	0.59	0.02	Hungary	0.65	0.01	1.00	0.03
					Bulgaria	0.68	0.01	1.00	0.03

It is reasonable to assume that in these countries most 15-year-olds and their parents would know whether they had ever repeated a grade. In contrast, it is quite reasonable that a fourth-grader might not readily know the number of books in his or her home. In both cases, plausible explanations for these findings are speculative at best; however, these two examples demonstrate a key problem that emerges time and again in international datasets—meaningful measurement error or misclassification is present in these variables. Furthermore, as the majority of information is collected from *either* the parent *or* the child (but infrequently from both), this issue usually takes the form of a missing-data problem. In both cases, missing and error-prone background questionnaires translate into biased subpopulation achievement estimates (Rutkowski, 2011, 2014; Rutkowski & Zhou, 2015), the degree to which depends on the missing and/or error mechanism. Importantly in the current example, a fairly straightforward question is posed to a relatively older population. It is reasonable to assume that more complex or subjective questions in younger populations will be even more error prone, giving rise to more meaningful problems in subpopulation achievement estimates.

It goes without saying that issues around grade repetition might figure differently into policy conversations than other key reporting variables such as immigrant status or SES. Nevertheless, it remains that self-reported data are error prone, and ignoring or failing to take account of this error produces undesirable analytic results and possibly misdirected policy interventions or initiatives. One possible solution relies on collecting data from multiple sources or from a source that can be regarded as *more* reliable. For example, school records could provide information on grade retention. Regarding sociocultural or SES, short, targeted questionnaires could be administered to the parents; however, this can add expense and logistical complexity to studies that are already ambitious in scale and scope.

Regardless, the policy and research importance of socioeconomic and sociocultural status cannot be underestimated. As a result, to the degree possible, future ILSA designs should include provision for better measures of key reporting variables that are also highly susceptible to measurement error. In economically well-developed educational systems where census-type data are collected (e.g., Norway or United States), reliable measures of school district SES can be derived using sophisticated approaches such as the U.S. Census's *small area income and poverty estimates* (SAIPEs; U.S. Census Bureau, 2016). Although this still leaves a gap between what we know about the school and the student, these sorts of measures are better and finer grained than anything used to date in international assessments. As an additional point, highly policy-relevant measures such as socioeconomic or sociocultural status will very likely be conceptualized and operationalized differently across educational systems. And although it is a reasonable goal to have a measure that is universal, this should not preclude individual countries from developing and including locally relevant measures of these sorts of variables to maximize the usefulness of international assessments. To that end, a short discussion of issues that are specific to measuring, especially, SES in international studies such as PISA, TIMSS, and PIRLS is included. Much of what follows draws on previously published work (Rutkowski & Rutkowski, 2013).

Measuring SES Internationally

Socioeconomic background typically relates to an individual's (or family's) status within a given social hierarchy. In a report on improving the measurement of SES in the U.S. National Assessment of Educational Progress (Cowan et al., 2012), the commissioned expert panel defines SES as

> one's access to financial, social, cultural, and human capital resources. Traditionally a student's SES has included, as components, parental educational attainment, parental occupational status, and household or family income, with appropriate adjustment for household or family composition. An expanded SES measure could include measures of additional household, neighborhood, and school resources. (p. 4)

In her extensive meta-analysis of SES research in education, Sirin (2005) notes that

> Regardless of disagreement about the conceptual meaning of SES, there seems to be an agreement on Duncan, Featherman, and Duncan's (1972) definition of the tripartite nature of SES that incorporates parental income, parental education, and parental occupation as the three main indicators of SES. (p. 418)

Sirin further explains that although research has demonstrated some correlation between these factors, "components of SES are unique" and should be "considered to be separate from the others" (p. 418). This three-factor approach to SES has also been found to explain achievement gaps better than a unidimensional approach (White, 1982). In the PISA study, the OECD has likewise taken a three-factor perspective on measuring socioeconomic background. However, as Hauser (2013) notes, the three components are combined into a single index of SES, muddling the unique contributions of the components to outcomes such as achievement.

Illustrating some of the measurement issues associated with this construct is the *wealth* index in PISA, which is one component of the *household possessions* index, taken together as a proxy for parental income. According to the PISA technical report (OECD, 2014b), this scale comprises eight international items asking students about their household possessions (a room of their own, a link to the Internet, a DVD player, and the number of cellular phones, televisions, computers, cars, and bath/shower rooms in their house). There is also the possibility of up to three country-specific items, such as a guest room, a high-speed Internet connection, or a musical instrument in the United States. Notably, in highly economically developed countries, some possession items suffer from low variance, adding little or no information to the scale. For example, 95 percent of Nordic participants (including Denmark, Finland, Iceland, Norway, and Sweden) answered yes to questions about a room of their own and an Internet connection. Furthermore, the OECD median reliability for this scale is 0.62, with a low of 0.53 in the Netherlands, indicating that there is nearly as much noise in the measure as actual signal. In contrast, the non-OECD median reliability is 0.74, suggesting that these items are more reliable measures of income in less economically developed countries. Along with the evidence of inconsistent responses between parents and children on the books in the home variable in PIRLS (a component of the SES measure in PISA), there is much work to be done to better measure SES internationally. And an important part of further efforts in this area is weighing the trade-offs of maximizing cross-

cultural comparability or within-country relevance of a highly relevant variable, such as SES. Finally, a comprehensive perspective on the issues associated with measuring SES internationally is outside the scope of the current paper; however, the above discussion serves to introduce and highlight the fact that it is an important area in need of in-depth research and strategies for improvement.

DESIGN-BASED CONSIDERATIONS FOR MAKING CAUSAL INFERENCES

Increasingly, international assessments are used as the basis for making causal inferences. Of course, as international assessments have grown in prominence as well as the number of studies and participants, a natural interest in understanding variation in achievement has developed in tandem. Perhaps more important, researchers and policy makers want to know what, if anything, can be done to improve achievement overall and for particular groups of test takers. This, in turn, has motivated interest in making connections between a host of potential causes and, generally although not exclusively, achievement. There is, however, a clear limitation in observational, cross-sectional studies such as TIMSS, PIRLS, and PISA—they do not meet the gold standard for making causal claims (e.g., via a randomized controlled trial, Meldrum, 2000) in "scientifically based research." Nevertheless, a collection of *quasi-experimental* methods exist that can be used to estimate "causal effects" from observational data. These methods rely on the early ideas of Hume and the counterfactual theory of causality (e.g., Rubin, 1974) or what is often referred to as the potential outcomes framework or the *Rubin causal model* (RCM; Holland, 1986). This causal approach emphasizes the *what-if* aspect of a sequence of events. What would we have observed if we could see the outcomes for one subject that had received *both* the treatment and the control? Of course, this is impossible in practice and is referred to as the *fundamental problem of causal inference* (Holland, 1986, p. 947). Rubin's causal model places emphasis on the *effects of the cause* and permits an estimate of the *average* causal effect of the treatment over a population of subjects. Importantly, observable information from different subjects can be used to inform us about the causal effect of the treatment.

Of critical importance in applying the RCM in observational settings such as those that are part and parcel of international assessments is the degree to which design choices permit a particular study to closely approximate randomized experiments (Rubin, 2007). Rubin's use of the words *study* and *design* in this context has a certain and perhaps not intuitive meaning. First, his description of a study in this case is a research question (e.g., the effect of school choice on achievement). In the context of international assessment, this is separate (but not completely so) from the larger study (e.g., TIMSS) that gives rise to the data used in a given study. And Rubin's use of the word *design* emphasizes the model used to statistically match the *treatment* group to the *control* group on important covariates that are known or believed to affect the treatment mechanism. If subject groups can reasonably be regarded as homogeneous on relevant covariates, average differences on the outcomes could be attributed to the treatment. Again, this consideration is somewhat, but not entirely, apart from the larger study. To that end, notions of *study* and *design* are inextricably linked to the characteristics of the extant dataset used to the answer a given research question. For example, in a study of school choice on achievement in the United States, race/ethnicity (Lubienski

& Lubienski, 2006) should be included as one of several covariates in a model to match treatment and control groups. In situations where this variable (or some reasonable proxy) is unavailable for inclusion, unobserved heterogeneity in the treatment variable is likely to produce biased estimates of the effect of school choice on achievement.

Rubin (2007) gauges our willingness to use a new drug, the safety and efficacy of which was evaluated by typical social science methods, to demonstrate that causal inferences should be the product of a carefully designed and executed study that is fit for answering the question at hand. The types of questions asked of ILSA data to answer causal questions have been observed firsthand in a recent special issue on causal inference with international assessment data (Rutkowski, 2016). In all papers in that issue, there were meaningful unanswerable questions regarding the degree to which critical assumptions of the methods used were met. Of course, these limitations are clearly delineated in the relevant section of the paper; however, it is important to explicitly recognize these limitations and to be vigilant about the impact that unmet assumptions can have on causal inferences and associated policy prescriptions. In the same issue, Rutkowski and Delandshere (2016) provided a useful framework for evaluating the tenability of causal inferences in ILSA settings. Using two prominent examples, the authors show that even in experimental studies, it is a real challenge to ensure that causal inferences are *valid* and that any conclusions are carefully scrutinized. To that end, Rutkowski and Delandshere (2016) note that the control required for making causal inferences necessitates a research question that is focused, qualified, and limited in scope (e.g., *Can a counseling intervention reduce dropout rates among at-risk populations?*). In contrast, policy makers are often interested in answers to broad questions (e.g., *How can we improve graduation rates among at-risk populations?*).

In support of the above argument, a review of the most recent TIMSS (International Association for the Evaluation of Educational Achievement, 2013) and PISA (OECD, 2013) science framework documents demonstrate that both studies are interested in science achievement in general. Certainly, interest in comparisons across educational systems disaggregated across select subpopulations is present. But there are no specific research questions posed by TIMSS or PISA study centers, and ancillary variables that are collected along with achievement measures are *generally* of interest to set the context of what students know and can do. Notably, what is measured by both studies is carefully developed and determined by panels of experts and agreed on by the consortium of participating education systems. So far, however, the study frameworks have not emphasized or identified *causal questions* to be answered. To do so, Kaplan (2016) recommends the development of a carefully defined set of causal questions that are integrated into the study framework. As an additional condition for asking causal questions of ILSA data, Kaplan also argues that along with a carefully developed treatment variable, it is also important to articulate (and operationalize) the context in which a cause occurs. These contexts are not causes in and of themselves; however, their consideration and measurement are important for *isolating* a given cause and estimating its effect. And as Rubin (2007) notes in his U.S. tobacco litigation example, identifying the important variables on which to match is not trivial and should be based on expert opinion. Many of the covariates in Rubin's example are biometric (e.g., diagnoses of high blood pressure and diabetes) or otherwise highly personal (e.g., public assistance status). Of course, ethical considerations

and perceptions of intrusiveness must be balanced against research interests in large cross-national studies. Note, however, that ensuring valid causal inferences will rely on the thoughtful development of causal questions as well as the important measures required to estimate causal effects.

A second option for strengthening the possibility to estimate causal effects from large-scale assessment data lies in adding a longitudinal or repeated-measures component to these studies. TIMSS is one study that could serve as a natural testbed for such an approach. Because fourth and eighth graders are assessed in TIMSS and the lag between measurements is 4 years, the fourth-grade population is randomly equivalent to the eighth-grade population 4 years later. Of course, there is the additional burden of tracking the longitudinal subset of the grade 4 sample, from, say, 2011 to 2015. A clear advantage in the TIMSS design is that there is not a need for an altogether new math or science test. However, a sufficient set of items should be developed that allows for linking across two tests, which could prove challenging given a 4-year gap in education. And although making claims about the effects of particular causes stands on a stronger foundation in a repeated measures study, a challenge remains that other plausible, intervening causes can be difficult to reject. Nevertheless, multiple measures over time on the same group of students would certainly be a move in the right direction when causal inferences are of interest. Furthermore, having such a repeated measures design would serve as a basis from which to analyze the effect of exogenous shocks, such as the recent economic crisis, on educational achievement among TIMSS-participating educational systems across cycles (e.g., between 2007 and 2011). Finally, any longitudinal component would need to be supplemented with the kind of careful development of causal questions as outlined in Kaplan (2016).

In closing this section, the tenuous nature of causal inferences with these data, particularly as they currently stand, is again highlighted. The cross-sectional, observational nature poses real challenges to convincingly inferring the effect of a cause. And although an assortment of quasi-causal designs and associated analytic methods exist, authoritatively concluding that all necessary assumptions are met is often beyond reach, given the restricted nature of available data. Given the current state of ILSAs, it is rather a more judicious approach to refer to estimates from procedures such as instrumental variables, propensity score matching, and other approaches as "less biased." Such nomenclature recognizes the limitations of the data while also acknowledging that care was taken to eliminate or minimize alternative explanations for observed effects.

CONCLUSION

In their present incarnation, ILSAs are the product of decades of careful methodological research, willingness on the part of stakeholders to engage in and support these massive endeavors, and a bit of trial and error. Through this process, ILSAs have evolved considerably in terms of the measured constructs and the populations of study participants. And the results have the potential to shine a light on the state of some aspects of an educational system at a given moment in time. These studies are also situated in a context of rapid global demographic changes, advances in technology, and changing stakes of international assessments. It is clear, too, that many recent ILSA innovations reflect a keen recognition of these changes (e.g., the adoption of

computerized testing platforms and modifications of tests and test content for different populations). Notwithstanding (and as with any major, high-profile undertaking), there is always room for further development and improvement. In accordance with the task of highlighting views on the design aspects of ILSAs that are most in need of revision, I outlined three areas, including issues around cultural comparability, the non-unique problem of measurement (or misclassification) error in survey research, and the fundamental challenge of drawing causal inferences with ILSA data. In each case, there are design considerations that could be applied to these issues. As possible solutions, further developments are in order that make ILSAs more relevant to individual participating educational systems.

Such solutions should take into consideration the specific cultural context of a country or region and directly incorporate them into the study design. Similarly, where key reporting variables figure prominently into policy discussions, decisions, and interventions, efforts should be made to reduce measurement or misclassification error to the degree possible. Where feasible, solutions that collect data from more reliable or more objective sources should be considered. Finally, as pressure to draw causal inferences from international assessment data mount, including a 2007 American Educational Research Association (AERA) report on the topic (Schneider et al., 2007) and the continued recommendation of AERA to consider this report when submitting applications to its research grant program, it is clear that researchers will continue to have an acute interest in the topic. As such, ILSA programs can integrate select causal questions into future study designs, offering the opportunity for such inferences to stand on a more principled foundation than currently allowed. Alternatively (or complementarily), a second option lies in including a repeated measures component to international assessments, with TIMSS providing the most natural place to further develop this idea.

Admittedly, none of these recommendations are simple or inexpensive to implement. Rather, each one requires adequate time and resources to design and evaluate particular solutions to the problems described. It is also reasonable that another scholar would highlight other problems or different solutions to the same problems, but as ILSAs grow in policy and research prominence, developmental and improvement efforts should be commensurate with the level of importance placed on these studies. As ILSAs are asked to do more and more (from system monitoring to providing the basis for causal inference), their long-term sustainability and credibility rely on providing valid, reliable evidence for fulfilling these lofty uses and interpretations. It is reasonably arguable, then, that such high-profile, cross-cultural, self-reported data would benefit from these or similar developments in future cycles.

REFERENCES

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185–198. http://doi.org/10.1111/j.1745-3984.1999.tb00553.x.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508. http://doi.org/10.1080/10705511.2014.919210.

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). International surveys of educational achievement: How robust are the findings? *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(3), 623–646. http://doi.org/10.1111/j.1467-985X.2006.00439.x.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. http://doi.org/10.1080/10705510701301834.

Cowan, C. D., Hauser, R. M., Kominski, R. A., Levin, H. M., Lucas, S. R., Morgan, S. L., Spencer, M. B., & Chapman, C. (2012). *Improving the measurement of socioeconomic status for the National Assessment of Educational Progress: A theoretical foundation*. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/researchcenter/Socioeconomic_Factors.pdf.

Dobbins, M., & Martens, K. (2012). Towards an education approach à la finlandaise?: French education policy after PISA. *Journal of Education Policy, 27*(1), 23–43. http://doi.org/10.1080/02680939.2011.622413.

Duncan, A. (2013, March 12). *The threat of educational stagnation and complacency*. U.S. Department of Education. Retrieved from http://www.ed.gov/news/speeches/threat-educational-stagnation-and-complacency.

Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). *Socioeconomic background and achievement* (Vol. 23). Oxford, England: Seminar Press.

Egelund, N. (2008). The value of international comparative studies of achievement: A Danish perspective. *Assessment in Education: Principles, Policy & Practice, 15*(3), 245–251. http://doi.org/10.1080/09695940802417400.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3–4), 199–215. http://doi.org/10.1080/15305058.2002.9669493.

Ercikan, K. (2003). Are the English and French versions' of the Third International Mathematics and Science Study administered in Canada comparable?: Effects of adaptations. *International Journal of Educational Policy, Research and Practice, 4*, 55–75.

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

Finn, Jr., C. E. (2010, December 8). A *Sputnik* moment for U.S. education. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052748704156304576003871654183998.html.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(3), 378–402. http://doi.org/10.1207/s15328007sem1303_3.

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

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

Grek, S. (2009). Governing by numbers: The PISA "effect" in Europe. *Journal of Education Policy, 24*(1), 23–37.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*(1), 69–86. http://doi.org/10.1016/j.stueduc.2007.01.006.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313–334. http://doi.org/10.1207/s15324818ame0204_4.

Hauser, R. M. (2013). Some methodological issues in cross-national educational research: Quality and equity in student achievement. *EurAmerica, 43*(4), 709–752.

He, J., & van de Vijver, F. J. (2013). Methodological issues in cross-cultural studies in educational psychology. In *Advancing cross-cultural perspectives on educational psychology.* Charlotte, NC: Information Age. Retrieved from http://www.researchgate.net/profile/Fons_Van_de_Vijver/publication/259176988_Methodological_Issues_in_Cross-Cultural_Studies_in_Educational_Psychology/links/02e7e52a2175522d1d000000.pdf.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945. http://doi.org/10.2307/2289064.

International Association for the Evaluation of Educational Achievement. (2013). *TIMSS 2015 assessment frameworks*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426. http://doi.org/10.1007/BF02291366.

Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-Scale Assessments in Education, 4*(1). http://doi.org/10.1186/s40536-016-0022-6.

Kreiner, S., & 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. http://doi.org/10.1007/s11336-013-9347-z.

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22*(2), 259–267.

Lubienski, S. T., & Lubienski, C. (2006). School sector and academic achievement: A multilevel analysis of NAEP mathematics data. *American Educational Research Journal, 43*(4), 651–698. http://doi.org/10.3102/00028312043004651.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mazzeo, J., & von Davier, M. (2009). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. Presented at the NCES Conference on the Program for International Student Assessment: What Can We Learn from PISA?, Washington, DC: IES National Center for Education Statistics.

Meldrum, M. L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America, 14*(4), 745–760. http://doi.org/10.1016/S0889-8588(05)70309-9.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational and Behavioral Statistics, 7*(2), 105–118. http://doi.org/10.3102/10769986007002105.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. http://doi.org/10.1007/BF02294825.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161. http://doi.org/10.3102/10769986017002131.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*(2), 131–154.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Boston: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Boston: TIMSS & PIRLS, International Study Center, Lynch School of Education, Boston College. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED512411.

OECD (Organisation for Economic Co-operation and Development). (2010). *TALIS technical report*. Paris: OECD Publishing.

OECD. (2013). *Draft PISA 2015 science framework*. Paris: OECD Publishing.

OECD. (2014a). *PISA 2012 results: What students know and can do*. Paris: OECD Publishing.

OECD. (2014b). *PISA 2012 technical report*. Paris: OECD Publishing.

OECD. (2014c). *TALIS 2013 technical report*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/edu/school/TALIS-technical-report-2013.pdf.

OECD. (n.d.). *PISA 2012 compendium for the cognitive item responses*. Paris: OECD Publishing. Retrieved from http://pisa2012.acer.edu.au/downloads/M_comp_COG_DEC03.zip.

Oliveri, M. E., & 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., & 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. http://doi.org/10.1080/15305058.2013.825265.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. http://doi.org/10.1037/h0037350.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. http://doi.org/10.1093/biomet/63.3.581.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.
Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*(1), 20–36. http://doi.org/10.1002/sim.2739.
Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education, 4*(1). http://doi.org/10.1186/s40536-016-0019-1.
Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259–278.
Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement, 48*(3), 293–312. http://doi.org/10.1111/j.1745-3984.2011.00144.x.
Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education, 27*(2), 115–132. http://doi.org/10.1080/08957347.2014.880440.
Rutkowski, L. (Ed.). (2016). Causal inferences with cross-sectional large scale assessment data. *Large-Scale Assessments in Education, 4*(8). http://doi.org/10.1186/s40536-016-0019-1.
Rutkowski, L., & 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. http://doi.org/10.1080/00220272.2010.487546.
Rutkowski, L., & Rutkowski, D. J. (2016). Improving the comparability of international assessments: A look back and a way forward. Manuscript submitted for publication.
Rutkowski, L., & Zhou, Y. (2015). The impact of missing and error-prone auxiliary information on sparse-matrix sub-population parameter estimates. *Methodology, 11*(3), 89–99. http://doi.org/10.1027/1614-2241/a000095.
Rutkowski, L., Rutkowski, D., & 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.
Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.
Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling* (Vol. 23). Oxford, UK: Ballinger.
Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. http://doi.org/10.3102/00346543075003417.
U.S. Census Bureau. (2016, March 8). *Small area income and poverty estimates.* Retrieved from http://www.census.gov/did/www/saipe/index.html.
von Davier, M. (2015). Notes on scaling, linking, fairness, comparability. Presented at the PISA International Research Conference, Oslo, Norway. Retrieved from https://www.berg-hansen.no/eventportal/?E=433&A=49509&Att=0&WebNo=1&Sec=jfdjcibAifAhkENS.
von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Mongraph Series, 2*, 9–36.
White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461–481. https://doi.org/10.1037/0033-2909.91.3.461.