International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting

Leslie Rutkowski, Eugenio Gonzalez, Marc Joncas, and Matthias von Davier

The technical complexities and sheer size of international large-scale assessment (LSA) databases often cause hesitation on the part of the applied researcher interested in analyzing them. Further, inappropriate choice or application of statistical methods is a common problem in applied research using these databases. This article serves as a primer for researchers on the issues and methods necessary for obtaining unbiased results from LSA data. The authors outline the issues surrounding the analysis and reporting of LSA data, with a particular focus on three prominent international surveys. In addition, they make recommendations targeted at applied researchers regarding best analysis and reporting practices when using these databases.

Keywords: assessment; international education/studies; statistics

In the past two decades, the use of international surveys of educational achievement for research has increased substantially. The two flagship studies of the International Association for the Evaluation of Educational Achievement (IEA)—the Progress in Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS)—have grown from 35 and 46 participating educational systems to 41 and 59 participating systems over their 8- and 15-year program histories, respectively (TIMSS & PIRLS International Study Center, n.d., 2008a, 2008b, 2009). An analogous study by the Organisation for Economic Co-operation and Development (OECD)—the Programme for International Student Assessment (PISA)—has grown from 43 participating educational systems in 2000 to 57 in the 2006 study (OECD, n.d.). An August 2009 search on the Wilson Education Full Text database for *TIMSS or PIRLS or PISA* indeed confirms a steady increase in research involving these databases. This search resulted in 340 articles published from 1995 to 1999; 556 articles published from 2000 to 2004; and 851 articles published over the period 2005 to 2009.

As these studies grow, the possible research questions and the data available for further investigation also greatly increase; however, the technical complexities and sheer size of international large-scale assessment (LSA) databases often cause hesitation on the part of researchers. As the specialized approaches to data analysis are generally not part of many university curricula outside survey methods programs or courses, the consequences of ignoring the complex data structure are often overlooked. In this article, we outline some of the main issues surrounding the analysis and reporting of LSA data, with a particular focus on international surveys including TIMSS, PIRLS, and PISA. Further, we make recommendations targeted at applied researchers regarding best analysis and reporting practices. We limit much of our discussion and all of our examples to IEA studies (TIMSS and PIRLS); however, the issues and implications outlined in this article are generally applicable to PISA and other LSA programs. We believe our contribution complements but differs in a meaningful way from the work of Thomas and Heck (2001) and Thomas, Heck, and Bauer (2005) in that we specifically target approaches for analyzing three commonly used data sets that demand additional considerations for achievement scores and, in the case of IEA studies, teacher-level analyses.

In addition to cross-sectional estimates of achievement, TIMSS, PIRLS, and PISA feature a wealth of contextual background information from participating students and their homes, teachers, and schools. Student background questionnaires solicit information regarding attitudes toward learning, home environment, study and leisure habits, and perceptions of school climate, among a host of other student background domains. Further, PIRLS and PISA include a questionnaire for parents or guardians regarding the home environment. At the teacher level, TIMSS and PIRLS include the collection of information on pedagogical practices, perceptions of teachers' own preparation, and school climate. Finally, all three studies include responses from principals to questions about school resources, climate, and other school-level issues surrounding teaching and learning. These databases provide an excellent resource for researchers interested in the contexts and correlates of learning internationally.

Why the Data Are Different and Implications for Analysis

International LSA data, like most survey data, differ in many ways from more traditional data sets. Broadly, LSA data (including international surveys) are not sampled at random, and each student is not administered every available cognitive item (Martin, Mullis, & Kennedy, 2007; Olson, Martin, & Mullis, 2008; OECD, 2009). This unique design necessitates the use of

Educational Researcher, Vol. 39, No. 2, pp. 142–151
DOI: 10.3102/0013189X10363170
© 2010 AERA. http://er.aera.net

sampling weights, particular variance estimation techniques, and a special method of analyzing achievement data, which is available through specialized software such as the IDB Analyzer (2009) or online tools such as the NAEP Data Explorer (National Center for Education Statistics, n.d.). Because of the sampling strategy used in IEA studies, a further consideration involves correctly interpreting findings when working with teacher data. We explain each of these issues in turn.

Sampling

All three assessment programs, PIRLS, PISA, and TIMSS, use a complex two-stage clustered sampling design (Martin et al., 2007; Olson et al., 2008; OECD, 2009). In Stage 1, schools are chosen based on a probability proportional to (the school's) size sampling approach, whereby larger schools are chosen with higher probability (see Brewer & Hanif, 1983, for a review of this approach). In IEA studies, the second stage consists of choosing randomly one or two intact classes at the fourth-grade (TIMSS and PIRLS) or eighth-grade (TIMSS) level. All students in selected classes are then assessed. Alternatively, the PISA approach results in the random selection of a set number of individual students (usually 35) from each sampled school's list of 15-year-olds. Every selected student is then administered a background questionnaire along with a subset of cognitive items. Given these sampling approaches, it should be evident that LSA studies do not follow a simple random sampling approach, whereby each student in the target population is chosen with equal probability. Therefore, standard variance formulas found in any introductory statistics text are not appropriate. Further, sampling weights, discussed in the next section, should be used for most analyses.

In IEA studies, where intact classrooms are assessed, the sampling framework poses an additional constraint for researchers. In particular, inferences at the teacher level are not appropriate because teachers are not explicitly sampled. Instead, students are sampled and their associated teachers are administered a questionnaire. Under this restriction, it is suitable to ask, Do students taught by teachers who are content with their profession experience more pedagogical approaches to reading than do students taught by teachers who are not content? Conversely, it is *not* appropriate to ask, Do teachers who are more content with their profession as a teacher employ more pedagogical approaches to reading than do teachers who are not content? Although this might appear a caviling point, the difference in interpretation is important—we can only generalize to students, not to their teachers. Practically speaking, any analysis that examines teacher attributes should employ teacher data that are merged to student-level data, and investigations should proceed as student-level analyses with teacher-level variables interpreted as student attributes.

Weights

In LSA studies, sampling weights are used to accommodate the fact that some units (schools, teachers, or students) are selected with differing probabilities. A simple example is used to illustrate. Assume that we are interested in the study habits of a particular classroom with 18 students, 12 boys and 6 girls. If we randomly choose 6 students to participate in our study, we would expect to select 4 boys and 2 girls on average. Assume, however, that it is important to include an equal number of boys and girls in our study while accounting for the fact that girls represent a smaller proportion of students in our hypothetical class. Using this approach, we would select 1 girl for every 2 students surveyed, giving each girl a 3/6 probability of selection. Similarly, we would also survey 1 boy for every 2 students selected, implying a 3/12 probability of selection for each boy because we would choose 3 of 12 possible boys. To ensure that girls are not overrepresented in our resulting estimates, every surveyed student's response is adjusted to reflect the student's actual proportional occurrence in the population. These adjustments are the sampling weights.

Raw (also called *unscaled*) sampling weights are derived in such a way that the sum of the weights within a sample adds up to the number of units in the population. Further, the inverse of the weight assigned to a unit (in our case a student) represents the probability of selection for that given unit. For example, if Student A has a weight of 10, Student A was selected with a probability of 1/10 = 0.10. Using the earlier example, each of our 3 hypothetical girls would carry a sampling weight of 2, whereas each of the boys would carry a sampling weight of 4. Summing the weights of the girls (2 + 2 + 2 = 6) and the boys (4 + 4 + 4 = 12) leads us back to our original population size of 18.

In LSA studies, ignoring sampling weights essentially gives more importance to some students, based solely on decisions linked to the choice of the sampling design. Thus, sampling weights should usually be used when conducting analyses with these data. Again, our hypothetical example illustrates the danger of ignoring weights. Assume the 3 selected girls report that they study 3, 2.25, and 2 hours per night and the 3 selected boys report that they study 1, 0.75, and 1.50 hours per night. The resultant unweighted mean would be estimated as (3 + 2.25 + 2 + 1 + 0.75 + 1.50)/6 = 1.75, whereas the weighted mean would be estimated as [2 × (3 + 2.25 + 2) + 4 × (1 + 0.75 + 1.50)]/18 = 1.53, which is nearly 13% smaller than the unweighted estimate.

To understand better the practical impact of using (or failing to use) weights, consider the results from TIMSS 2007 Bulgarian eighth-grade mathematics achievement data: The unweighted mean is 481.38, whereas the weighted mean (as reported in the international report) is 463.63. The observed difference was the result of the Bulgarian sampling design that gave a higher probability of selection to students from vocational and profiled schools (Olson et al., 2008). Average student achievement from those schools was nearly 80 points higher than student achievement from general schools. Ignoring the sampling weights essentially allowed for a disproportional contribution from students who attended vocational and profiled schools. In our example, the unweighted achievement estimate would have placed Bulgaria at position 18 (of 49 participating countries) in the mathematics achievement rankings rather than its correct position of 23. These are markedly different results with possibly important policy implications, particularly when achievement is compared among participating countries. Further, if unweighted data are used, researchers might be unnecessarily alarmed to find that their results differ from those of published study reports.

Here we detail the most important weights and the types of analyses to which they should be applied. In TIMSS and PIRLS, there are generally five important sets of weights: (a) total student weight, (b) student house weight, (c) student senate weight,

(d) overall and subjectwise teacher weight, and (e) school weight. Sets (b), (c), and (d) all are linear transformations of (a) that result in some number with desirable properties. Total student weight is appropriate for single-level student-level analyses in survey software such as SUDAAN (2001), AM (AM Statistical Software, 2009), WesVar (2007), SAS SURVEYMEANS (SAS, 2003), and SPSS COMPLEX SAMPLES (SPSS, 2007), all of which adjust standard errors to reflect the study design. But the limited or specialized functionality of these applications often necessitates the use of other software and, thus, different sets of weights.

In analyses that are especially sensitive to sample size (e.g., chi-square tests), we can instead use normalized weights, called *house weights* in TIMSS and PIRLS. House weight is essentially a linear transformation of total student weight so that the sum of the weights is equal to the sample size. Note that linear transformations of the total student weight will have no effect on resultant point estimates, such as means. For example, if an individual has a sampling weight of 10 and another individual has a sampling weight of 5, this has the same effect on estimates as if those same individuals had weights of 4 and 2, respectively. In both instances, the first unit has a weight twice as large the second unit. Consider another example. The number of U.S. fourth-grade students surveyed for PIRLS 2006 was 5,190. Thus, if we sum the house weights for each sampled U.S. student, we will obtain 5,190. If we instead sum the total student weight for each sampled student, we arrive at 3,351,959—the approximate size of the population of fourth-grade students in the United States.

For analyses that combine more than one country, it is important to consider the differential impact that proportionally larger countries can have on results. For example, the size of the U.S. eighth-grade population is estimated at more than 3 million students, whereas the eighth-grade population of Israel is just 84,000. An analysis that compares the effect of gender on achievement across these two countries using total student weight would be dominated by the U.S. solution. To adjust for different population sizes, TIMSS and PIRLS include *senate weight* in the database for cross-country analyses. Senate weight[1] is student total weight scaled in such a way that all students' senate weights sum to 500 in each country. In analyses that include more than one country, senate weight ensures that each country contributes approximately equally.[2] As an example, we performed a maximum likelihood factor analysis for Egypt ($N_{\text{total}} \approx 1$ million) and the United States on four variables that constitute the TIMSS index Positive Affect Towards Math. The student total and senate weighted solutions are located in Table 1. When total student weight is used, there is a stronger relationship between the four individual variables and the latent variable *positive affect.* In the senate-weighted solution, we see an attenuated relationship between the four individual variables and the latent construct, suggesting that the United States was more influential in the exploratory factor solution because of its proportionally larger population.

For analyses that include teacher variables as attributes of the student, it is appropriate to use the overall teacher weight (for all PIRLS analyses and for TIMSS analyses that include both science and math teachers). If teachers of a particular subject (either math or science) in TIMSS are of analytic interest, it is then suitable to select either math teacher weight for analyses that include math teacher variables or science teacher weight for analyses that include science teacher variables. Bear in mind that these weights are just student total weight divided by the total (or subject-specific) number of teachers a student has. These are nonnormalized weights, and as such, they might influence the results of sample-size-sensitive analyses.

Table 1
Total and Senate Weighted Factor Analysis Solutions Using TIMSS 2007 Data

	Unrotated Factor Loadings (i.e., Structure Coefficients)	
	Total Weight	**Senate Weight**
I usually do well in math	0.74	0.64
Math is more difficult for me	–0.62	–0.54
Math is not one of my strengths	–0.73	–0.68
I learn things quickly in math	0.71	0.64

Finally, if a researcher is interested in school-level attributes, both TIMSS and PIRLS databases contain a school-level weight that is the inverse of the probability of selection for the selected school. As with total student weight, the school weight is a raw weight and sums to the size of the target population—in this case, an estimate of the total number of schools in a given country.

So far, the discussion on weight selection has been limited to single-level analyses; however, the clustered structure of the data often demands a multilevel approach. Weighting survey data for multilevel analyses is not a straightforward issue, and researchers interested in a multilevel approach should carefully consider the level of analysis when selecting weights. Programs such as HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004) and Mplus (Muthén & Muthén, 1998–2007) allow users to specify weights at multiple levels. Thus, in a two-level model, it is possible to select both Level 1 and Level 2 weights. Analysts might be tempted to assign the readily available school weight at Level 2 and total student weight at Level 1. We advise against this practice, as there is an assumption that the Level 1 student weight (when Level 2 weights are included) is inversely proportional to the probability of a student being selected *given* that the school was selected (Chantala & Suchindran, 2006; Raudenbush et al., 2004); however, the total student weight is essentially the inverse of the *joint* probability of selection for a particular student (the probability that School A *and* Class C *and* Student B are selected).

Multilevel analysts should consult their software documentation for the appropriate application of weights at multiple levels. In particular, issues of weight scaling and parameter estimation are important considerations, and approaches differ (e.g., Asparouhov & Muthén, 2006; Chantala & Suchindran, 2006; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998; Rabe-Hesketh & Skrondal, 2006). Regardless of the particular software, applying weights at multiple levels requires that users manually calculate weights for each level. TIMSS and PIRLS databases include the individual pieces of information necessary to decompose total student weight into the inverse of the probability of selection for a school, for a class within a school given that the associated school was selected, and for a student given that the student's school and class were selected. We first discuss the following formula for the overall student sampling weight for student i ($i = 1, \ldots, n_j$) in class j ($j = 1, \ldots, n_k$) in school k ($k = 1, \ldots, K$), where n_j is the number of students in class j, n_k is the number of classes in school k, and K is the total number of schools in the sample:

$$\textit{Total Student Weight}_{ijk} = \underbrace{(WF_k \times WA_k)}_{\frac{1}{\substack{\text{adjusted}\\ \text{probability of}\\ \text{selection for}\\ \text{school } k}}} \times \underbrace{(WF_j \times WA_j)}_{\frac{1}{\substack{\text{adjusted}\\ \text{conditional}\\ \text{probability of}\\ \text{selection for}\\ \text{class } j}}} \times \underbrace{(WF_i \times WA_i)}_{\frac{1}{\substack{\text{adjusted}\\ \text{conditional}\\ \text{probability of}\\ \text{selection for}\\ \text{student } i}}}$$

WF is defined as the weight factor or inverse of the probability of selection for the relevant unit (school, class, or student); *WA* is a weight adjustment or nonresponse adjustment for units that were sampled but did not participate.[3] As indicated in the formula, the product of the school weight factor (WF_k) and weight adjustment (WA_k) gives the inverse of the probability of selection for school k adjusted for schools that refused to participate. In the case of classes, the product ($WF_j \times WA_j$) gives the inverse of the conditional probability of selection for class j given that school k was selected for participation. Finally, the student-level product ($WF_i \times WA_i$) is the inverse of the probability that student i is selected given that that student's class was selected. In line with IEA documentation, we refer to each of these weights as *final weights.*

Consider an example that uses one student's data from PIRLS 2006. The selected student was assigned a total student weight of 22.71. This implies that the student's probability of selection was 1/22.71 or 0.04. What follows is a decomposition of our exemplar student's total weight into its component parts:

$$\begin{aligned} 22.71 &= \underbrace{(14.11 \times 1.00)}_{\frac{1}{\substack{\text{adjusted}\\ \text{probability of}\\ \text{selection for}\\ \text{school } k}}} \times \underbrace{(1.50 \times 1.02)}_{\frac{1}{\substack{\text{adjusted}\\ \text{conditional}\\ \text{probability of}\\ \text{selection for}\\ \text{class } j}}} \times \underbrace{(1.00 \times 1.05)}_{\frac{1}{\substack{\text{adjusted}\\ \text{conditional}\\ \text{probability of}\\ \text{selection for}\\ \text{student } i}}} \\ &= (14.11) \times (1.53) \times (1.05). \end{aligned}$$

The probability of selection for this student's school was 1/(14.11 × 1.00) = 0.07, whereas the probability of selection for this student's class was 1/(1.50 × 1.02) = 0.65. Finally, given that this student's school and class were selected, the probability of selection for this student was 1/(1.00 × 1.05) = 0.95.

For the analyst interested in a three-level model that accounts for variance at the school, class, and student levels, it is necessary to generate the *final student weight,* the *final class weight,* and the *final school weight* and apply the relevant weights at each of the levels prior to analysis. For our example student, final student weight = (1.00 × 1.05) = 1.05; final class weight = (1.50 × 1.02) = 1.53; and final school weight = (14.11 × 1.00) = 14.11. In general, however, one class per school was selected for TIMSS and PIRLS. In this case, it is then advisable to combine the *final class weight* with the *final student weight* via multiplication to represent the inverse of the probability of selection for the student given that the school the student is in was selected. In our example, this weight is 1.53 × 1.05 = 1.61.

For two-level analyses that use schools as Level 1 and countries as Level 2, school weight at Level 1 is suitable. As an aside, two- (or three-) level analyses that use teacher data at Level 1 are not appropriate because of the sampling framework used in TIMSS and PIRLS. Although we have not covered every contingency for selecting and deriving weights, we have outlined strategies for many common situations. Finally, it is important to emphasize that the user must carefully consider the purpose of the analysis and the research question when selecting the sampling weights to be used.

Proficiency Estimation

In both IEA studies and PISA, a complex scheme of item administration is used to minimize testing burden on individual students while ensuring that accurate population estimates of proficiency can be generated from the available data. We briefly outline these methods here, using TIMSS 2007 as the exemplar; however, the methods used across all three studies are generally similar and originate from methods developed for the National Assessment of Educational Progress (Beaton & Johnson, 1992; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992; von Davier, Sinharay, Oranje, & Beaton, 2006). For details on the particular assessment designs and proficiency estimation procedures, consult the relevant study's technical documentation and assessment framework (Martin et al., 2007; Mullis, Kennedy, Martin, & Sainsbury, 2006; Mullis et al., 2005; Olson et al., 2008; OECD, 2005, 2009). To ensure that items receive sufficient exposure in the sample and that enough items are administered to individual students to estimate population proficiency reliably, a complex rotated booklet design was used. Specifically, items are assembled into a nonoverlapping set of blocks with 10 to 15 items per block. In 2007, the eighth-grade TIMSS assessment included 429 total mathematics and science items distributed across 14 mathematics blocks (M01–M14) and 14 science blocks (S01–S14), and arranged into 14 booklets with 4 blocks each. Under this design, each block (and therefore each item) appears in two booklets. These 28 blocks of items represent more than 10 hours of testing time; however, the booklet design used by TIMSS reduced individual testing time to 90 minutes per student plus 30 minutes for the student background questionnaire. A representation of the TIMSS 2007 assessment design can be seen in Table 2.

Although this administration method minimizes testing time for students, it poses challenges for generating individual proficiency estimates. In particular, traditional methods of estimating individual proficiency result in biased or inconsistent variance estimates of population parameters (Mislevy, Beaton, et al., 1992; von Davier, Gonzalez, & Mislevy, 2009). Instead, plausible value methods (Mislevy, 1991) are employed as a viable technique for generating *population-level* proficiency estimates from test designs where only a small number of items from the total item pool are administered to any given student. Essentially, under the plausible value methodology, student achievement is treated as a missing value. Responses to the limited subset of administered cognitive items and complete student background questionnaires

Table 2
Trends in International Mathematics and Science Study 2007 Fourth- and Eighth-Grade Student Achievement Booklet Design

Student Achievement Booklet	Part 1		Part 2	
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

Note. M = mathematics block; S = science block.

are used in conjunction with a measurement model–based extension of Rubin's (1987) multiple imputation approach to generate a student ability distribution for the population (or subpopulation) of interest (Beaton & Johnson, 1992; Mislevy, Beaton, et al., 1992; Mislevy, Johnson, et al., 1992; von Davier et al., 2006). From the estimated ability distribution, several random draws, each referred to as a plausible value, are selected for every student. That is, a fixed number (usually five) of individual proficiency estimates are drawn at random from an empirically derived ability distribution. In practice, five plausible values are reported for each student in overall math and science (TIMSS and PISA) and five for each student in overall reading (PISA and PIRLS). In addition, five plausible values are present for each of the subdomains assessed. In TIMSS 2007, for example, students receive five values for each of the science subdomains Chemistry, Earth Science, Biology, and Physics.

These estimates are then used to calculate consistent estimates of population characteristics for the whole country or for subpopulations of policy interest, such as mathematics achievement for girls of highly educated parents in a given country. There is an important distinction between this approach and others where an emphasis is placed on generating ability estimates for *individual* examinees and the objective of the assessment is to obtain individual scores for program selection, accountability, or other individual examinee–focused testing purposes. The technical details surrounding ability estimation in the international LSA context are beyond the scope of the present article (von Davier et al., 2009; von Davier et al., 2006). Instead, we note the practical implications of plausible value methods for analysts of LSA data.

From a practical standpoint, treating the additional values is a straightforward matter: Standard analyses are performed on each of the plausible values, and the results of each analysis are combined. For instance, if a researcher is interested in estimating the relationship between time spent on homework and science achievement, a regression model of science achievement on time spent on homework is fit five times—once with each of the plausible values. The results of these five analyses are then combined into a single set of point estimates and standard errors using formulas following Rubin's (1987) multiple imputation approach. See Schafer's (1999) article for an accessible primer on multiple imputation and detailed instructions for combining parameter estimates to achieve unbiased results.

When conducting analyses that use plausible values, two *shortcuts* are often implemented, both of which are incorrect. First (and less problematic of the two methods), analysts often choose to use just one of the five plausible values, say the first overall reading score in PIRLS. When analyzing only one plausible value, the standard errors of the statistics of interest are generally underestimated, as the uncertainty associated with the measurement of proficiency distributions is ignored. We illustrate with an example from Latvia. We initially regressed each of the plausible values onto a measure of time spent watching television and then combined each of the parameter estimates according to Rubin (1987) for comparison. It is important to note that in these analyses we did not consider the error associated with sampling on standard error estimates. Thus these estimates did not account for the sampling variance, and they are underestimates. We treat the issue of variance estimation in the next section.

Results can be found in Table 3. Of interest are the first five estimates of the effect of time spent watching television (PV1, PV2, . . . , PV5) compared with the last estimate (combined). Here, we can see that the point estimates across the plausible values are quite different, whereas the standard errors are very close to one another. Also apparent is the potential for contradictory findings, depending on which plausible value is chosen. For three out of five plausible values, we find a statistically significant relationship between time watching television and mathematics achievement in Latvia; whereas in two instances, the relationship was not statistically significant. How do we reconcile these differences? Which one is the best to report? When dealing with plausible values, the best estimate is the estimate that accounts for the imputation variance and averages the point estimates across plausible values for the effect of television exposure. The additional imputation variance is reflected in the larger standard error (and statistically nonsignificant *p* value) associated with the combined estimate.

The second shortcut often employed by researchers is to use the mean of the five plausible values as a single estimate of achievement for individuals. Averaging plausible values in this manner results in standard errors that are even more severely underestimated than are estimates using a single set of plausible values. Consequently, averages of plausible values should never be used in analyses (von Davier et al., 2009). As can be seen in Table 3, the point estimate for the effect of time watching television (TV) on the average of the reading plausible values for Latvia (labeled *PV avg.*) is equivalent to the combined estimate that replicated the analysis across all five plausible values. However, it can be seen that the standard error of the coefficient for television exposure is even smaller than in any of the analyses that used single plausible values and certainly smaller than in the analyses that account for the uncertainty associated with imputing student ability. This smaller standard error might result in different conclusions, depending on the a priori accepted Type I error rate.

Table 3

Estimates and Standard Errors for Progress in Reading Literacy Study 2006 Analyses That Use Single, Average, and Combined Plausible Values (PV)

Variable	Coefficient	*SE*	*t*	*df*[a]	*p* value	R^2
TV (PV1)	−1.93	0.91	−2.12	4,112	.03	.01
TV (PV2)	−1.89	0.92	−2.05	4,112	.04	.01
TV (PV3)	−1.56	0.92	−1.71	4,112	.09	.01
TV (PV4)	−0.90	0.92	−0.98	4,112	.33	.01
TV (PV5)	−2.12	0.92	−2.30	4,112	.02	.01
TV (PV average)	−1.68	0.86	−1.94	4,112	.05	.01
TV (combined)	−1.68	1.06	−1.59	67	.12	.01

Note. Only variability due to imputation, not sampling, is represented in the standard errors reported here. TV = effect of time watching television.
[a] *The much larger degrees of freedom (df = 4,112) result from a standard ordinary least squares regression on each of the individual and averaged plausible values, whereas the reduced degrees of freedom for the combined analysis (df = 67) result from necessary adjustments due to combining five estimates from each of the plausible values, according to Rubin (1987).*

In our experience, we have also found that there is a temptation on the part of various stakeholders to use results based on plausible values as evidence for incentive schemes for schools, teachers, and students. This practice tends to manifest itself in the form of rank-ordered students (or classes) based on the mean of the plausible values for each student or higher level unit. Again, it is important to reiterate that the methods used to generate proficiency estimates are useful only for making population-level inferences and not for making inferences for small groups. Briefly, we illustrate with an example of the variability across plausible values within a given unit. From a randomly chosen student, Student R, we calculated a number of statistics based on the weighted means of each of the five PIRLS 2006 reading scale plausible values for her class, her school, and her country. These are compared with Student R's plausible values and the mean of her plausible values.

Moving from left to right across the Table 4, from the country-level plausible value means to the individual plausible values for Student R, a number of points are worth noting. Although the plausible value mean for Student R was higher than the national average, the variability associated with the plausible value means at the class, school, and country were markedly lower than for the individual student. That is, the uncertainty associated with group-level estimates is much less than for student-level estimates. The same was true of the range in values of the mean versus the individual plausible values—from a low of 535.19 to a high of 591.90 in the case of Student R, more than a 56-point difference. These statistics suggest an intolerably high degree of uncertainty surrounding an individual's set of plausible values. All of the variance estimates in this example ignored the uncertainty associated with sampling and are therefore underestimates. We discuss this topic in the next section.

From a strictly pragmatic standpoint, the barriers to reporting at the student level should be obvious: Imagine the outcry that would ensue if five different, highly varied values were reported to individual students and their parents. Among the myriad questions, "Which value is *true?*" would likely top the list. With these barriers in mind, analysts are reminded that only reporting at the subpopulation level is appropriate (e.g., reporting by gender, parental education, or immigration background groups). Reporting at the individual student level is never appropriate.

Finally, we urge caution on the part of researchers who might have identifying information, not normally available to the public, that allows them to link outside data to the databases provided by the relevant study center. Examples of this sort of external information can include (but are not limited to) data on community demographics such as population concentrations, infrastructure, or the existence of after-school programs. Any information that is added to the database ex post facto is not used in the plausible value estimation and might bias estimates that use the newly added data to an unknown degree. This is analogous to a mismatch between the imputer's model and the analyst's model in the multiple imputation literature (Schafer, 2003).

Variance Estimation

Because of the stratified multistage sampling design used by most international LSAs, discussed earlier, the simple random sampling assumptions for calculating standard errors of estimates do not apply, and therefore the standard variance formulas for parameter estimates are not appropriate. Thus it is necessary for analysts to employ special methods for estimating unbiased measures of the uncertainty associated with sampling. Although a number of approaches exist, both the IEA and the OECD use replication methods for variance estimation. In particular, TIMSS and PIRLS employ the jackknife repeated replication (JRR) method, and PISA uses a modified balanced repeated replication (BRR) method. We recommend that the interested reader consult Rust (1985) for a discussion and comparison of these and other variance estimation methods; Johnson and Rust (1992) for details on the JRR as it was applied to TIMSS and PIRLS; and Judkins (1990) for details on Fay's modification of the BRR as it was applied to PISA.

Technical details aside, it is important for the analyst to recognize that standard variance estimation formulae are not appropriate for TIMSS, PIRLS, or PISA data. Instead, where applicable, JRR or BRR methods should be used for all sampling variability estimates (i.e., standard errors). Readily available variance estimation applications are generally limited to means, correlations, multiple linear and nonlinear regression, and categorical measures of association. Although many of these procedures are available in SPSS as add-on modules, as part of the base SAS platform,

Table 4
Progress in Reading Literacy Study 2006 Plausible Value (PV) Means at Several Levels of Analysis for a Randomly Chosen Student, Student R

	Student R's Country	Student R's School	Student R's Class	Student R
READING (PV1)	521.50	527.73	534.13	535.19
READING (PV2)	521.44	527.01	531.00	565.73
READING (PV3)	521.46	530.25	542.43	589.18
READING (PV4)	521.92	530.91	534.21	591.90
READING (PV5)	521.32	528.61	531.26	572.52
PV means	521.53	528.90	534.61	570.90
PV mean minimum	521.32	527.01	531.00	535.19
PV mean maximum	521.92	530.91	542.43	591.90
PV mean range	0.60	3.89	11.43	56.71
PV mean variance	0.05	2.71	21.44	519.82

and in freely available software (IDB Analyzer, 2009; WesVar, 2007), procedures for generating unbiased variance estimates might be unfamiliar for those who are new to the analysis of LSA data. Thus it is recommended that researchers familiarize themselves with these methods via assessment program user guides (Foy & Kennedy, 2008; Foy & Olson, 2009; OECD, 2005) and manuals for the software that handles complex surveys.

We illustrate the importance of calculating unbiased measures of sampling variability via our earlier example of the association between student television exposure and TIMSS 2007 mathematics achievement for Latvian students. According to Table 3, our best estimate of the variability of the television effect was 1.06; however, this estimate included only the variability associated with imputing student ability in the form of five plausible values. To obtain a better estimate of the total variability—both sampling and imputation—we reestimated the regression of mathematics achievement on time spent watching television, using the freely available survey software IDB Analyzer (2009). We found an identical point estimate for the television exposure effect (–1.68); however, the standard error that incorporates both components of variability was much larger (1.66). This variance value is a more realistic reflection of the uncertainty associated with the estimated relationship. It is nearly twice the earlier estimate that used the mean of the plausible values and more than 50% larger than the estimate that accounts for the imputation variance.

Survey Data and Causality

All three studies discussed herein are cross-sectional surveys. That is, different cohorts of students are surveyed at each time point on each study. Further, TIMSS, PIRLS, and PISA are observational studies without random assignment of students to treatments. Therefore, language that implies causality should generally not be used when discussing the results of analyses of these studies. In particular, terms such as *factors that influence, variables that impact, changes in* x *cause* y, and so on, are not appropriate. Of course, this brief list does not exhaustively account for all possible causal language. It is instead appropriate to use correlational language such as *variables associated with, factors that predict,* or *a relationship exists between* x *and* y.

Summary

In this article, we briefly explained a number of design issues that set international LSA data apart from more traditional data that result from experimental or randomly sampled observational settings. We also discussed a number of pitfalls that can lead to incorrect results and interpretations when researchers analyze international LSA data such as TIMSS, PIRLS, and PISA. We illustrated the issues at hand using examples, and we discussed the consequences of ignoring a number of factors that necessitate careful treatment of this type of data. Although our discussion was not exhaustive, we covered the most prominent issues, including weights, variance estimation, dealing with plausible values, and level of analysis. In the last section, we briefly conclude with a summarized checklist of issues for researchers to consider when analyzing international LSA data such as TIMSS, PIRLS, or PISA.

Summary of Recommendations for Principled Analysis and Reporting

This section briefly outlines a general approach for researchers who are using TIMSS, PIRLS, or PISA data to answer their research questions. Although we target these three assessments, the recommended approach is generally applicable to most survey data with the possible omission of proficiency value considerations when dealing with surveys where every respondent receives every item. The following are a set of general steps that researchers should consider when planning analyses involving the studies discussed in the current article.

1. With a clear research question in mind, carefully choose an appropriate sampling weight. Weights should generally be applied to every analysis. This is important for both point estimates and standard error estimates.
2. When proficiency estimates are of interest, plausible values should be used in conjunction with survey software for general analyses, including calculating means and standard errors and single-level linear and logistic regression analyses. For methods not covered by survey software, analyses should be repeated five times, once with each of the plausible

Table 5
Resources for Researchers

Area	Resources	Location
TIMSS and PIRLS	General study information	http://timss.bc.edu/
All IEA studies	Documentation, data, SPSS and SAS scripts, and free analyzer	http://www.iea.nl/iea_studies_datasets.html
LSA in general	IEA–Educational Testing Service Research Institute (IERI)—Training opportunities and open source for IERI Monograph	http://www.ierinstitute.org/
OECD PISA	Study information, documentation, SAS and SPSS scripts, and data	http://www.pisa.oecd.org

values. The resulting parameter estimates should be combined using Rubin's (1987) guidelines to account for imputation variability associated with generating plausible values.

3. Resampling variance estimation techniques, available for standard analyses supported by survey software, should be used whenever possible to ensure unbiased estimates of sampling variability. When such variance estimation techniques are not supported for a particular method, it is important to report that the standard errors might be underestimated.
4. When conducting analyses that involve teacher-level data, always analyze and report teacher variables as attributes of the student. In multilevel models that involve teacher data, students should always be the lowest level of analysis.
5. Finally, consider that TIMSS, PIRLS, and PISA are cross-sectional and observational studies. Because of this design, it is generally not appropriate to make causal inferences using these data.

SAS and SPSS scripts that correctly perform many of the most common analyses (generating means, correlations, and linear regressions) are available in the relevant database user guides and on each study's associated website. Further, the IEA makes available the IDB Analyzer (2009), a free, user-friendly plug-in for SPSS that performs the same analyses as do the SPSS scripts. Finally, we recommend that researchers not rely solely on this article or the documentation provided by the assessment programs. Instead, we encourage researchers to take advantage of training sessions offered during research conferences such as the American Educational Research Association annual meetings (http://www.aera.net/AnnualMeeting.htm) or by research organizations conducting these assessments. Table 5 provides information on how to locate these resources.

With advances in methods and software, we expect that applications suitable for analyzing survey-type data will grow. It is possible that methodological advances will render less useful the recommendations contained in the current article. For example, a number of software packages now readily handle plausible values and expand the off-the-shelf applications available to analysts. Two examples include PROC MI ANALYZE in SAS (SAS, 2003) and HLM 6 (Raudenbush et al., 2004). Further, there currently exist innovative multilevel applications that are sufficient to address the complex design and the requisite variance estimation (Stapleton, 2008). In the present article, we have outlined the broad issues associated with analyzing international LSA data and the considerations necessary for principled analysis and reporting. Applied researchers should view this article as a primer on the issues and methods important for generating unbiased results from LSA data. We encourage researchers to acquaint themselves with these interesting and useful databases and methods. As familiarity and expertise with issues covered in this article increase among researchers, we anticipate that the results of international LSAs can be more fully exploited to benefit education domestically as well as internationally.

To conclude, we ask the question that might be on many reader's minds: Why deal with such complex data in the first place? We offer a few answers to this question. The databases that are one of the end products of international LSAs are the result of a rigorous, collaborative process that provides a wealth of comparable data on the context of learning internationally. Those interested in topics such as motivation, school or student resources, and achievement; methodological issues surrounding survey data, item response theory, or sampling; and international comparisons in any of the aforementioned areas will find these data sets a good starting point to develop or validate theories in many areas of social science, including (but certainly not limited to) education, economics, and psychology.

NOTES

We would like to thank three anonymous reviewers, the editorial team at *Educational Researcher,* and our colleagues at Educational Testing Service and the IEA Data Processing and Research Center for their constructive feedback on this article.

[1] *House weight* and *senate weight* are so called because the function of each weight is similar to that of its namesake in the U.S. Congress (the number of representatives in the House of Representatives is based on each state's population size, whereas the Senate gives two seats to each state regardless of population).

[2] We say "approximately" because this will ultimately depend on the amount of missing data for the variables in question.

[3] In the International Association for the Evaluation of Educational Achievement databases, these weight factors and weight adjustments have the variable names WGTFAC$_m$ and WGTADJ$_m$, where m = 1, 2, or 3 for school, class, and student, respectively.

REFERENCES

AM Statistical Software [Manual]. (2009). Washington, DC: American Institutes for Research.

Asparouhov, T., & Muthén, B. (2006). Multilevel modelling of complex survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association 2006*, 2718–2726.

Beaton, A., & Johnson, E. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement*, *29*, 163–175.

Brewer, K., & Hanif, M. (1983). *Sampling with unequal probabilities, lecture notes in statistics* (*Vol. 15*). New York: Springer-Verlag.

Chantala, K., & Suchindran, C. (2006). Adjusting for unequal selection probability in multilevel models: A comparison of software packages. *Proceedings of the Survey Research Methods Section, American Statistical Association 2006*, 2815–2824.

Foy, P., & Kennedy, A. (2008). *PIRLS 2006 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, P., & Olson, J. (Eds.). (2009). *TIMSS 2007 international database and user guide*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

IDB Analyzer (Version 2) [Computer software and manual]. (2009). Hamburg, Germany: International Association for the Evaluation of Educational Achievement.

Johnson, E., & Rust, K. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*, *17*, 175–190.

Judkins, D. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, *6*, 223–239.

Martin, M., Mullis, I., & Kennedy, A. (Eds.). (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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

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

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154.

Mullis, I., Kennedy, A., Martin, M., & Sainsbury, M. (2006). *PIRLS assessment framework and specifications*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I., Martin, M., Ruddock, G., O'Sullivan, C., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthén, L., & Muthén, B. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

National Center for Education Statistics. (n.d.). NAEP Data Explorer. *In National Center for Education Statistics*. Retrieved January 7, 2010, from http://nces.ed.gov/nationsreportcard/naepdata/

Olson, J., Martin, M., & Mullis, I. (Eds.). (2008). TIMSS 2007 technical report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Organisation for Economic Co-operation and Development. (n.d.). *Participating countries*. Retrieved December 31, 2009, from http:http://www.pisa.oecd.org/pages/0,3417,en_32252351_32236225_1_1_1_1_1,00.html

Organisation for Economic Co-operation and Development. (2005). *PISA 2003 data analysis manual*. Paris: Author.

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris: Author.

Pfeffermann, C., Skinner, D., Holmes, H., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *60*, 23–40.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society. Series A (General)*, *169*, 805–827.

Raudenbush, S., Bryk, A., Cheong, Y., & Congdon, R. (2004). *HLM 6 [Manual]*. Lincolnwood, IL: Scientific Software International.

Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. New York: John Wiley.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, *1*, 381–397.

SAS (Version 9) [Computer software and manual]. (2003). Cary, NC: SAS Institute.

Schafer, J. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3–15.

Schafer, J. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*, 19–35.

SPSS (Version 16.0) [Computer software and manual]. (2007). Chicago: SPSS, Inc.

Stapleton, L. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling*, *15*, 183–210.

SUDAAN (Release 8.0) [Manual]. (2001). Research Triangle Park, NC: Research Triangle Institute.

Thomas, S., & Heck, R. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education*, *42*, 517–540.

Thomas, S., Heck, R., & Bauer, K. (2005). Weighting and adjusting for design effects in secondary data analysis. *New Directions for Institutional Research*, *2005*(127), 51–72.

TIMSS & PIRLS International Study Center. (n.d.). *Countries*. Retrieved December 29, 2009, from http://timss.bc.edu/pirls2001i/pirls2001_countries.html

TIMSS & PIRLS International Study Center. (2008a). *Countries participating*. Retrieved December 29, 2009, from http://timss.bc.edu/pirls2006/countries.html

TIMSS & PIRLS International Study Center. (2008b). *Countries participating*. Retrieved December 29, 2009, from http://timss.bc.edu/timss2003i/countries.html

TIMSS & PIRLS International Study Center. (2009). *Countries participating*. Retrieved December 29, 2009, from http://timss.bc.edu/TIMSS2007/countries.html

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9–36.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam: Elsevier.

WesVar 4.3 [Manual]. (2007). Rockville, MD: Westat.

AUTHORS

LESLIE RUTKOWSKI is an assistant professor of inquiry methodology in the Department of Counseling and Educational Psychology at Indiana University, 201 North Rose Avenue, Bloomington, IN 47405; *lrutkows@indiana.edu*. Her research focuses on international large-scale assessment from both a methodological and applied perspective, and her interests include the impact of background questionnaires on assessment results and integrating survey methods with statistical modeling to better exploit large-scale assessment data.

EUGENIO GONZALEZ is a program administrator at Educational Testing Service and unit head of the Research and Analysis Unit at the IEA Data Processing and Research Center, 666 Rosedale/Carter Road, Princeton, NJ 08540; *egonzalez@ets.org*. His research interests include design and implementation of large-scale assessments and psychometrics.

MARC JONCAS is senior methodologist in the Social Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, R. H. Coats Building, Floor 15 G, Ottawa ON K1A 0T6; *Marc.Joncas@statcan.gc.ca*. His areas of expertise in statistics include topics such as complex survey design, survey data weighting, complex survey computational statistics, variance estimation, complex survey statistical analysis (nonlinear regression, chi-square, survival analysis), and the analytical methodologies used in large-scale educational surveys such as PISA, TIMSS, PIRLS, and IALL.

MATTHIAS VON DAVIER is principal research scientist in the Research and Development Division, Educational Testing Service, MS 02-T, Princeton, NJ 08541; *mvondavier@ets.org*. His areas of expertise and research in psychometrics include topics such as item response theory; latent class analysis; classification and mixture distribution models; diagnostic models; computational statistics; person-fit, item-fit, and model checking; hierarchical extension of models for categorical data analysis; and the analytical methodologies used in large-scale educational surveys such as PISA, TIMSS, PIRLS, PIAAC, and NAEP.

Manuscript received May 20, 2009
Final revisions received January 11, 2010
Accepted January 19, 2010