

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation.* Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 29 Number 4, March 2024 ISSN 1531-7714

|  |
| --- |
| **Statistical Analyses with Sampling Weights in Large-Scale Assessment and Survey** |

Ting Shen, *Missouri University of Science and Technology*

|  |
| --- |
| Large-scale assessment and survey (LSAS) studies have been increasingly utilized to address important research questions. As LSAS data adopt multi-stage complex sampling designs with unequal probabilities of selection, sampling weights need to be used. However, it is unclear how to use different sampling weights variables in statistical analysis using LSAS data. Thus far, research evidence and practical guidelines have been scarce and inconsistent. Using data from Canada, Italy, and Lithuania in PISA and TIMSS that represent two main sampling frameworks in LSAS, this study examined unweighted and weighted analyses and compared weighted single-level versus multi-level models with different sampling weights variables (e.g., student weights, house weights, senate weights, replicate weights, and school weights). Findings reveal that weighted estimates were different from unweighted estimates, and statistical significance of model analyses may vary for some variables in different countries and different LSAS data. For weighted multilevel models, three approaches (i.e., size scaling, effective scaling, school weights only) generated more similar results. Practical recommendations are provided.Keywords: complex sampling weights, statistical analysis, large-scale assessment and survey |

Introduction

 The Organization for Economic Cooperation and Development (OECD) and the International Association for the Evaluation of Educational Achievement (IEA) have invested tremendous resources to conduct large-scale assessment and survey (LSAS) studies. The Program for International Student Assessment (PISA) and the Trends for International Mathematics and Science Study (TIMSS) are two respective flagship international studies that have been conducted for more than two decades. One advantage of LSAS studies is that they provide reliable and comparable assessment measures of achievement data plus rich contextual background information about students’ homes, teachers, and schools. Another advantage of LSAS studies is that complex sampling designs are adopted to generate national probability samples that represent well-defined populations (e.g., 8th graders in the U.S.).

 However, the technical complexity of multistage sampling designs and different weights variables prevent applied researchers from utilizing these high-quality data to inform policy and practice in social sciences (Rutkowski et al., 2010). Thus far, there is no consensus about whether sampling weights should be used and how to use them in different statistical analyses. In LSAS, it is not uncommon that sampling weights were not used in model analysis even for papers published in top-tier journals (Osborne, 2011), or if used, information was not provided about what sampling weights variables were used and how they were used (Laukaityte & Wiberg, 2018). Notably, if sampling weights were not used or used inappropriately, the sample may not reflect the population due to complex sampling designs, so the estimates of population quantities could be biased. In addition, standard errors could be under- or over-estimated, so statistical inferences and findings of variables of interest might be different. Using sampling weights could appropriately reflect population characteristics and with appropriate variance estimation techniques, researchers could obtain unbiased estimates of population parameters (Stapleton, 2013). Therefore, applying sampling weights in statistical analysis is essential in LSAS data (Rutkowski et al., 2010).

 Nevertheless, how to use sampling weights in statistical analysis in LSAS is not straightforward. For example, it is unclear how to choose appropriate sampling weights variables (e.g., student weights, house weights, senate weights, and replicate weights) for conducting weighted statistical analyses. It is unknown whether researchers should use weighted single-level or weighted multi-level models with LSAS data (Koziol et al., 2017). Overall, very few studies have examined different approaches to employing weights variables in statistical analyses using LSAS. Research that examines weighted and unweighted approaches across multiple LSAS data with different sampling designs has not been performed. Therefore, providing more empirical evidence and guidelines about conducting weighted analyses especially weighted single-level and weighted multi-level models in the context of LSAS will inform and advance this field.

 To fill the literature gaps, this study examines how to apply complex weights in statistical analysis using the STATA software program and the PISA and TIMSS data that represent two main sampling frameworks in LSAS. In addition to comparing descriptive statistics with different weights variables, this study compares single-level versus multi-level models and single-country versus multi-country analyses using different weights variables (e.g., student weights, house weights, senate weights, replicate weights, and school weights). This study clarifies some controversies and summarizes prior research findings of incorporating complex sampling weights in statistical analysis, demonstrates how to use appropriate weights in statistical analysis in the context of LSAS, and provides practical recommendations. With more research evidence, knowledge, and practical guidelines, applied researchers will gain familiarity, confidence, and flexibility in choosing appropriate weights in conducting different statistical analyses.

Literature Review

 LSAS studies typically adopt a two-stage stratified cluster sampling framework with unequal probability sampling designs at school level. There are two conventional methods to approximate complex sample design estimators and to reduce bias in variance estimation: Taylor series linearization, and replication or resampling procedure which includes Jackknife repeated replication (JRR), balanced repeated replication (BRR), and bootstrap (Rust, 2013). In LSAS, replication methods are preferred to the linearization approach because to estimate the sampling variance, the linearization method needs a different variance formula for each type of estimator whereas a common formula can be used for a particular replication approach (Rust, 2013). Among the replication methods, JRR and BRR are more efficient than the bootstrap in most practical applications and thus are used commonly in LSAS (Rust, 2013).

 There are two limitations for using variance estimation methods to deal with multi-stage sampling designs. First, variation estimation methods are limited to simple estimation (e.g., means, correlations) and statistical models (e.g., multiple linear regression, nonlinear regression, and categorical analysis of associations), but if researchers are interested in complex statistical models (e.g., multilevel models), appropriate variance estimation methods are not available. Second, although these procedures are available in conventional software programs (e.g., SPSS, and SAS) and in the IDB analyzer provided by LSAS data, empirical researchers may not know how their software handle complex survey designs using variance estimation methods with replicate weights. In addition, recommendations are not consistent. For example, Stapleton (2013) recommended that researchers should use replicate weights in variance estimation procedures and should not use them in single-level or multilevel model analysis (Stapleton, 2013). Nevertheless, other researcher indicated that sampling weights could be used with replicate weights (Rutkowski et al., 2010).

 With respect to statistical modeling, single-level models and multi-level models were typically compared under equal probability sampling designs; however, under unequal probability sampling designs, little is known about differences between weighted single-level models and weighted multi-level models (Koziol et al., 2017). Thus far, one prior study compared weighted single-level and multilevel models, and found that they both have pros and cons (Koziol et al., 2017). Specifically, if sample conditions are not ideal, weighted single-level estimators are better than weighted multilevel estimators; when the outcome is not normally distributed or the true covariance structure is complex, weighted single-level modeling is questionable. However, LSAS studies recommend that sampling weights should be applied in all statistical analyses, but there is no guidance on when and how to implement weighted single-level or multilevel models using LSAS data.

 Practically, it is easier to utilize weighted single-level models with overall sampling weights to produce population estimates. One advantage of weighted single-level modeling is that it flexibly adjusts standard errors due to design features using survey variance estimation methods at the top level of primary sampling units. In contrast, although weighted multilevel modeling with level-specific weights is more complex in estimation, it matches with hierarchical data structures of multi-stage sampling; accordingly, one advantage of using multilevel modeling is that it controls for the design effect of clustering (Raudenbush & Bryk, 2002). Technically, in single-level modeling, the pseudo maximum likelihood (PML) has been utilized to deal with unequal probabilities of selection and produce consistent estimates (Binder, 1983). Although the PML works well for single-level models, it is not applicable to multilevel models where observations are dependent within clusters. Instead, in multilevel modeling, two general estimation methods have been proposed: the probability-weighted iterative generalized least squares (PWIGLS) (Goldstein, 1986; Pfeffermann et al., 1998) and the multilevel pseudo maximum likelihood (MPML) (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006).

 Overall, research evidence and practical guidelines for using sampling weights in statistical modeling are limited and inconsistent especially for weighted multilevel models in LSAS data. Carle (2009) provided a summary of findings from prior simulation work regarding incorporating design weights in multilevel models (Carle, 2009). One main finding is that scaling methods applied to level-one (i.e., lower level) weights provide better estimates than both the unweighted approach and the raw weights without scaling. In addition, there is no gold standard for scaling methods, as the results from two main scaling methods (i.e., size scaling and effectives scaling) depend on different factors such as the number of clusters and design weights’ informativeness. For example, a simulation showed that effective scaling outperformed size scaling in multilevel structural equation models (Stapleton, 2002) whereas other researchers found that size scaling was preferred (Pfeffermann et al., 1998).

 In addition, some recommendations from prior studies are not applicable to LSAS data. One study is used to illustrate this point. Asparouhov (2006) provided six-step procedure to avoid pitfalls of weighted multilevel modeling. The first step is to verify if sampling weights are designed for a single-level analysis or a multilevel analysis. If only single-level sampling weights were provided, then multi-level weights could not be obtained. This is a good general suggestion, however, in LSAS studies, both single-level sampling weights and multilevel sampling weights are provided, and both weighted single-level analysis and weighted multilevel analysis could be conducted. In the second step, Asparouhov (2006) recommended that if only cluster sampling weights are available, weighted multilevel models should be not used, because using cluster level weights only in multilevel models would generate more estimation bias than unweighted models (Asparouhov, 2006). However, using PISA data, Mang and colleagues (2021) found that using only cluster level weights in multilevel modeling is a favorable approach (Mang et al., 2021). In the sixth step, Asparouhov (2006) recommended that researchers could complete an informativeness test to determine if they can ignore sampling weights and use an unweighted analysis instead. However, this recommendation of possibly not using sampling weights is not consistent with the guidelines provided in LSAS user manuals, which indicate that researchers need to apply sampling weights in conducting statistical analysis using LSAS data. In addition, there is no specific guidance on conducting informative tests for weighted multilevel models in the context of LSAS data.

 As LSAS data (e.g., PISA and TIMSS) have a clustering data feature, multilevel models could simultaneously estimate individual level and group level predictors as well as provide correct standard errors of regression coefficients (Raudenbush & Bryk, 2002). If researchers are interested in conducting weighted multilevel modeling using PISA data, Mang et al. (2021) recommended to use three approaches: only school weights, weighted multilevel models with size scaling, and weighted multilevel models with effective scaling; they found that the two scaling approaches provide nearly identical results (Mang et al., 2021). When analyzing TIMSS data for multilevel modeling, Laukaityte and Wiberg (2018) found that in empirical analysis, using no weights or only level 1 (student level) weights may lead to misleading conclusions, but simulation analysis showed small differences in statistical significance between weighted and unweighted estimation under informative designs although the weighted unscaled approach had significant differences in estimating some parameters (Laukaityte & Wiberg, 2018). In general, for multilevel modeling, in a two-level model, student weights and school weights could be used, but it is not appropriate to use these weights variables as is (Rutkowski et al., 2010).

Research Method

**LSAS Data and Sampling Weights**

 This study uses two LSAS data (i.e., PISA and TIMSS). They represent two common sampling designs in LSAS data, and two different types of population: age-based (i.e., 15-year old students in PISA) and grade-based (4th or 8th grade in TIMSS) (Rust, 2013). The PISA study collects data on assessments of 15-year-old students’ achievement in mathematics, science, and reading literacy plus information about students’ learning environments and their educational experiences and attitudes via student, teacher, and principal surveys. In general, the sampling strategy in PISA involves a two-stage stratified sampling design where the first-stage sampling stage selected individual schools that had targeted 15-year-old students with probabilities proportional to size (PPS), and the second sampling stage sampled a fixed number of students with equal probabilities within sampled schools (OECD, 2014). PISA 2015 data provided school weights that included school base weights and adjustments for nonresponse, final student weights that included student base weights and school base weights with adjustments for nonresponse at school and student levels, and senate weights. Senate weights are student total weights scaled in a way that all students’ senate weights sum to an arbitrary number, which could be different across different LSAS data, so that each country has an equal contribution in statistical analysis. If analyses involve more than one country, researchers could use senate weights (Rutkowski et al., 2010). In addition, there are 80 BRR replicate weights.

The TIMSS study collects assessment data of 4th and 8th graders’ mathematics and science achievement as well as rich survey information from students, teachers, and school principals since 1995. TIMSS 2015 adopted a two-stage stratified cluster sampling design in which schools were primary sampling units with the PPS sampling technique, and within schools, one or more intact classrooms were sampled in which all students in a classroom were included in the sample (Olson et al., 2008). The TIMSS 2015 provided four sets of sampling weights variables: (1) student overall weights; (2) school overall weights; (3) student house weights; (4) student senate weights. In addition, JRR weights variables were provided.

The student overall weights are appropriate to be used in single-level analyses with some survey software such as SUDAAN, SAS SURVEYMEANS, and SPSS COMPLEX SAMPLES to adjust standard errors due to study sampling designs (Rutkowski et al., 2010). If the research interest is on school-level variables, overall school weights should be used. Both student overall weights and school overall weights are used for analysis within a country. If student-level analyses involve multiple countries, house weights or senate weights should be used to ensure equal weights or proportional weights for countries of different sample sizes when their data are combined (Rutkowski et al., 2010). House weight is a linear transformation of total student weights, so that the sum of the house weights would be equal to the total sample size, while the sum of total student weights would be approximately the population size (Rutkowski et al., 2010; Stapleton, 2013). In addition, if analyses are sensitive to sample size (e.g., chi-square tests), it is also appropriate to use house weights (Rutkowski et al., 2010). All these four sets of weights allow researchers to run weighted single-level analyses. Stapleton (2013) mentioned that if analysis is conducted within a single country, using student overall weights, senate weights and house weights would have same point estimates, but it is unknown if standard errors would be similar or not (Stapleton, 2013).

Statistical Analyses

 The statistical analysis involves three components to compare weighted and unweighted results, which include descriptive statistics, single-level models, and multilevel models. Data from three countries (i.e., Canada, Italy, Lithuania) and a combined data set are used for both PISA 2015 and TIMSS 2015. In terms of variables of interest, in PISA, the outcome is science score and four predictors include class size, a dummy variable of female student, school size, and percent of female students. In TIMSS 8th grade data, variables of interest include science score, student age, a dummy variable of female student, instructional hours per year, and percent of female students.

 For descriptive statistics, means and standard deviations of five variables are compared for four conditions in PISA (i.e., unweighted, total student weights, senate weights, and total student weights with BRR replicate weights) and five conditions in TIMSS (i.e., unweighted, total student weights, senate weights, house weights, and total student weights with JRR replicate weights).

With respect to model analysis, the single-level model for *i*th student can be written as:

 $Y\_{i}=β\_{0}+XB+e\_{i}$ , (1)

where *Y* is the outcome of science achievement, $β\_{0}$ is the intercept, $X$ refers to a row vector of four covariates including class size, a dummy variable of female student, school size, and percent of female students in PISA, whereas in TIMSS, the four covariates include student age, a dummy variable of female student, instructional hours per year, and percent of female students, ***B*** represents a column vector of corresponding coefficients, e is the error term assumed to follow a normal distribution with the mean zero and variance $σ\_{e}^{2}$. In the PISA data, in addition to the unweighted model analysis, three types of weights are used, which include total student weights, senate weights, and total student weights plus BRR weights. In TIMSS data, four types of weights variables are used including total student weights, senate weights, house weights, and total student weights with JRR weights; the results are then compared with the unweighted approach.

 For multilevel modeling, a two-level random intercept model is written as:

 $Y\_{ij}=β\_{0}+XB+u\_{0j}+ε\_{ij} , $ (2)

where$Y\_{ij}$ is the outcome of science achievement for student *i* in school *j*, $β\_{0}$ and $X$***B*** are same as in Equation (1), $u\_{0j}$ is the student residuals, and $ε\_{ij}$ is the school residuals. Comparing Equations (1) and (2), one difference is that the residual e in the single-level model is partitioned into two error terms in the multi-level model.

 There are six multilevel estimators: (1) unweighted two-level model, (2) multilevel model with student-level specific weights, (3) multilevel model with school weights, (4) multilevel model with student-level specific weights and school-level weights without scaling, (5) multilevel model with student-level specific weights plus size-scaling and school-level weights, (6) multilevel model with student-level weights plus effective scaling and school-level weights.

Results

 Table 1 displays the descriptive statistics of mean, standard deviation (SD), and sample size (N) for PISA 2015. The unweighted N represents the sample size whereas the weighted ones with total student weights and with senate weights represent the population size and the common size of 5000, respectively. Comparing weighted and unweighted descriptive statistics, overall, the means and SDs varied across variables of interest. For example, in Canada and Lithuania, the weighted means of science score, class size, and school size were larger than the unweighted ones. The weighted SDs of science score and school size were larger than the unweighted SDs, whereas the weighted SDs of class size were slightly smaller than unweighted ones. In Italy, the unweighted mean of science score was larger than the weighted means, whereas the unweighted means of class size and school size were smaller than weighted ones. In combined data, unweighted mean of science score was larger than the mean with student total weights, while the latter is slightly larger than the one with senate weights. However, for female and percent female, results with and without weights were consistent across four data sets. Regarding weighted descriptive statistics, within each of the three countries data, the means and the SDs of five variables were

identical using the weights variables, but for the combined data of the three countries, descriptive statistics of science score, school size, and percent female were slightly different using total student weights versus senate weights.

**Table 1.** PISA 2015 Weighted and Unweighted Descriptive Statistics

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Unweighted | Total Student Weights | Senate Weights |
|   | Mean | SD | Mean | SD | Mean | SD |
| **Canada** |   |   |   |   |   |   |
| Science score | 516.46 | 91.62 | 527.69 | 92.72 | 527.69 | 92.72 |
| Class size | 3.48 | 0.93 | 3.63 | 0.92 | 3.63 | 0.92 |
| Female | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| School Size | 772.36 | 421.79 | 922.80 | 441.60 | 922.80 | 441.60 |
| Percent female | 0.50 | 0.11 | 0.50 | 0.11 | 0.50 | 0.11 |
| N | 20,058 |   | 331,546 |   | 5,000 |   |
|   |   |   |   |   |   |   |
| **Italy** |   |   |   |   |   |   |
| Science score | 492.47 | 89.29 | 480.70 | 91.77 | 480.70 | 91.77 |
| Class size | 3.00 | 1.12 | 3.11 | 1.06 | 3.11 | 1.06 |
| Female | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| School Size | 760.77 | 352.89 | 833.96 | 358.52 | 833.96 | 358.52 |
| Percent female | 0.50 | 0.25 | 0.50 | 0.24 | 0.50 | 0.24 |
| N | 11,583 |   | 495,093 |   | 5,000 |   |
|   |   |   |   |   |   |   |
| **Lithuania** |   |   |   |   |   |   |
| Science score | 468.94 | 90.29 | 475.73 | 90.17 | 475.73 | 90.17 |
| Class size | 3.10 | 1.07 | 3.23 | 1.05 | 3.23 | 1.05 |
| Female | 0.49 | 0.50 | 0.49 | 0.50 | 0.49 | 0.50 |
| School Size | 513.86 | 256.26 | 533.20 | 259.99 | 533.20 | 259.99 |
| Percent female | 0.49 | 0.12 | 0.49 | 0.12 | 0.49 | 0.12 |
| N | 6,525 |   | 29,915 |   | 5,000 |   |
|   |   |   |   |   |   |   |
| **All three countries** |   |   |   |   |   |
| Science score | 501.05 | 92.45 | 498.71 | 94.92 | 494.70 | 94.50 |
| Class size | 3.27 | 1.04 | 3.32 | 1.04 | 3.32 | 1.04 |
| Female | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| School Size | 724.65 | 389.49 | 857.84 | 397.21 | 763.32 | 397.70 |
| Percent female | 0.50 | 0.17 | 0.50 | 0.20 | 0.50 | 0.17 |
| N | 38,166 |   | 856,554 |   | 15,000 |   |
| *Note.* BRR=Balanced repeated replication  |   |   |   |

**Table 2**. TIMSS 2015 (Grade 8) Weighted and Unweighted Descriptive Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Unweighted | Student Weights | Senate Weights | House Weights |
|   | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| **Canada** |   |   |   |   |   |   |   |   |
| Science score | 523.27 | 71.73 | 517.95 | 71.87 | 511.75 | 76.34 | 517.95 | 71.87 |
| Age | 13.98 | 0.47 | 13.96 | 0.48 | 14.15 | 0.60 | 13.96 | 0.48 |
| Female | 0.51 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| Instruction hours per year | 944.63 | 78.72 | 948.76 | 79.53 | 945.64 | 129.79 | 948.76 | 79.53 |
| Percent female | 0.50 | 0.10 | 0.50 | 0.09 | 0.50 | 0.08 | 0.50 | 0.09 |
| N | 8,757 |   | 229,175 |   | 500 |   | 8,757 |   |
|   |   |   |   |   |   |   |   |   |
| **Italy** |   |   |   |   |   |   |   |   |
| Science score | 508.63 | 76.60 | 504.44 | 77.22 | 517.95 | 71.87 | 504.44 | 77.22 |
| Age | 13.80 | 0.49 | 13.81 | 0.50 | 13.96 | 0.48 | 13.81 | 0.50 |
| Female | 0.50 | 0.50 | 0.49 | 0.50 | 0.50 | 0.50 | 0.49 | 0.50 |
| Instruction hours per year | 1029.65 | 118.69 | 1031.72 | 120.56 | 948.76 | 79.53 | 1031.72 | 120.56 |
| Percent female | 0.50 | 0.07 | 0.50 | 0.07 | 0.50 | 0.09 | 0.50 | 0.07 |
| N | 4,481 |   | 528,839 |   | 500 |   | 4,481 |   |
|   |   |   |   |   |   |   |   |   |
| **Lithuania** |   |   |   |   |   |   |   |   |
| Science score | 501.05 | 82.47 | 512.84 | 79.14 | 504.44 | 77.22 | 512.84 | 79.14 |
| Age | 14.63 | 0.42 | 14.69 | 0.38 | 13.81 | 0.50 | 14.69 | 0.38 |
| Female | 0.48 | 0.50 | 0.50 | 0.50 | 0.49 | 0.50 | 0.50 | 0.50 |
| Instruction hours per year | 861.81 | 109.54 | 856.44 | 119.59 | 1031.72 | 120.56 | 856.44 | 119.59 |
| Percent female | 0.50 | 0.07 | 0.50 | 0.07 | 0.50 | 0.07 | 0.50 | 0.07 |
| N | 4,347 |   | 27,263 |   | 500 |   | 4,347 |   |
|   |   |   |   |   |   |   |   |   |
| **All three countries** |   |   |   |   |   |   |   |   |
| Science score | 514.05 | 76.35 | 508.68 | 76.01 | 512.84 | 79.14 | 513.25 | 75.30 |
| Age | 14.09 | 0.56 | 13.88 | 0.52 | 14.69 | 0.38 | 14.10 | 0.58 |
| Female | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| Instruction hours per year | 945.82 | 114.79 | 1001.42 | 119.53 | 856.44 | 119.59 | 947.08 | 119.35 |
| Percent female | 0.50 | 0.09 | 0.50 | 0.07 | 0.50 | 0.07 | 0.50 | 0.08 |
| N | 17,585 |   | 785,277 |   | 1,500 |   | 17,585 |   |
| *Note.* JRR=Jackknife repeated replication  |   |   |   |   |   |   |

 Table 2 reports the descriptive statistics in TIMSS 2015 Grade 8. The results of N with house weights are same as the unweighted approach, whereas the ones with total student weights represent the population size, and the ones with senate weights sum to the common size of 500. Comparing the estimates of unweighted and various weighted means and SDs, there is no clear pattern for differences across five variables. For example, in Canada, the weighted means of science score were smaller than the unweighted mean, whereas in Lithuania, the weighted means of science score were larger than the unweighted mean. In Italy, the unweighted mean of science score was between the weighted mean with total weights and the weighted mean with senate weights. However, without formal statistical tests of mean differences, it is unknown whether the differences among different approaches were statistically significant.

 Table 3 shows the results of regression estimates, R-squares, and sample sizes for weighted and unweighted single-level models across four data in PISA 2015. Overall, three weighted approaches produced same regression coefficients and R-squares, which were different from the unweighted approach. The approaches of using total student weights and senate weights generated the same standard errors, which were slightly different from the approach of using both total student weights and BRR weights. The statistical significance varied across different variables in different datasets. In Canada, statistical significance was different only for female and percent female, comparing four model approaches. In Italy, the results of statistical significance across these four model approaches were consistent for three out of four variables. For class size, the approach of using total student weights plus BRR weights generated a non-significant estimate, whereas other three approaches of no weights, total student weights, and senate weights had significant estimates. In Lithuania, the results of statistical significance were different only for the variable female. For the combined data, statistical inferences were different for the variable percent female.

 Table 4 shows results of regression estimates, R-squares, and sample sizes for weighted and unweighted single-level models across four data in TIMSS 2015 Grade 8. Similar patterns were observed. Four weighted approaches produced same regression coefficients and R-squares, which were different from the unweighted approach. The standard errors generated from three weighted approaches (i.e., total student weights, senate weights, and house weights) were similar but were different from the approach of using both total student weights and JRR weights. Statistical inferences varied across different variables in different countries. In Canada and Italy, statistical significance was different for instruction hours per year. In Lithuania, statistical significance was different for female. In the combined data, for percent female and instruction hours per year, unweighted and weighted approaches had different statistical significance.

 Table 5 displays results of weighted and unweighted analyses in multilevel modeling in PISA 2015. Examining the intraclass correlation (ICC) values, the estimates of regression coefficients, and statistical significance across these six approaches, it was found that estimates of unweighted and student specific weights approaches were more similar. Among the weighted estimates, size scaling and effective scaling approaches had similar results, which were close to results of using school weights only. For example, in Canada, for class size and female, unweighted and student specific weights approaches had significant estimates at 0.1 level, whereas the estimates from other four weighted approaches were not significant. For school size, unweighted and student specific weights had significant estimates at 0.01, whereas school weights, size scaling, and effective scaling approaches generated significant estimates at 0.05 level and the unscaled estimate approach had a significant result at 0.1 level. In Lithuania, for female, the estimates of school weights, size scaling, and effective scaling were not significant, but the unweighted and unscaled estimates were significant at 0.1 level and the estimate of student specific weights was significant at 0.01 level. If using the combined data, for school size and percent female, unweighted and student specific weights had significant results whereas school weights, unscaled, size scaling and effective scaling had non-significant results.

 Table 6 shows results of weighted and unweighted analyses in multilevel modeling in TIMSS 2015 Grade 8. Overall, statistical significance was more consistent in TIMSS data than in PISA data. In Italy and in the combined data, for all four variables, the statistical significance was same whereas in Canada and Lithuania, only the variable instruction hours per year had different significant results across these six approaches. By and large, unweighted and student specific weights were similar. Comparing weighted approaches, size scaling and effective scaling generated similar estimates, which were closer to the approach of school weights. Unscaled approach was different from other approaches. For example, in Canada, for instruction hours per year, unweighted and student specific weights had significant estimates at 0.01 level, whereas the estimates from other four weighted

**Table 3.** PISA 2015 Weighted and Unweighted Single-Level (Student Level) Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Unweighted | Student Weights | Senate Weights | Student Weights +BRR Weights |
| **Canada** |   |   |   |   |
| Intercept | 481.416\*\*\* | 481.556\*\*\* | 481.556\*\*\* | 481.556\*\*\* |
|   | (3.871) | (5.940) | (5.940) | (8.350) |
| Class Size | 3.423\*\*\* | 6.172\*\*\* | 6.172\*\*\* | 6.172\*\*\* |
|   | (0.806) | (1.113) | (1.113) | (1.412) |
| Female | -2.215\* | -2.076 | -2.076 | -2.076\*\* |
|   | (1.313) | (1.866) | (1.866) | (0.911) |
| School Size | 0.025\*\*\* | 0.018\*\*\* | 0.018\*\*\* | 0.018\*\*\* |
|   | (0.002) | (0.002) | (0.002) | (0.003) |
| Percent female | 9.611 | 15.696\* | 15.696\* | 15.696 |
|   | (6.068) | (9.302) | (9.302) | (11.418) |
| R-Squared | 0.019 | 0.017 | 0.017 | 0.017 |
| N | 20058 | 331,546 | 5,000 | 331,546 |
| **Italy** |   |   |   |   |
| Intercept | 488.372\*\*\* | 450.705\*\*\* | 450.705\*\*\* | 450.705\*\*\* |
|   | (3.101) | (5.211) | (5.211) | (7.646) |
| Class Size | -2.235\*\*\* | 2.966\*\* | 2.966\*\* | 2.966 |
|   | (0.757) | (1.356) | (1.356) | (2.190) |
| Female | -17.573\*\*\* | -17.830\*\*\* | -17.830\*\*\* | -17.830\*\*\* |
|   | (1.910) | (2.840) | (2.840) | (1.077) |
| School Size | 0.025\*\*\* | 0.034\*\*\* | 0.034\*\*\* | 0.034\*\*\* |
|   | (0.002) | (0.004) | (0.004) | (0.005) |
| Percent female | 1.762 | 2.785 | 2.785 | 2.785 |
|   | (3.759) | (5.634) | (5.634) | (7.652) |
| R-Squared | 0.019 | 0.029 | 0.029 | 0.029 |
| N | 11583 | 495,093 | 5,000 | 495,093 |
| **Lithuania** |   |   |   |   |
| Intercept | 327.544\*\*\* | 325.096\*\*\* | 325.096\*\*\* | 325.096\*\*\* |
|   | (5.001) | (6.002) | (6.002) | (5.621) |
| Class Size | 17.527\*\*\* | 16.281\*\*\* | 16.281\*\*\* | 16.281\*\*\* |
|   | (1.141) | (1.355) | (1.355) | (2.050) |
| Female | -3.314 | -4.342\* | -4.342\* | -4.342\*\*\* |
|   | (2.139) | (2.358) | (2.358) | (1.111) |
| School Size | 0.048\*\*\* | 0.044\*\*\* | 0.044\*\*\* | 0.044\*\*\* |
|   | (0.005) | (0.005) | (0.005) | (0.006) |
| Percent female | 130.606\*\*\* | 156.322\*\*\* | 156.322\*\*\* | 156.322\*\*\* |
|   | (9.169) | (10.915) | (10.915) | (9.106) |
| R-Squared | 0.137 | 0.143 | 0.143 | 0.143 |
| N | 6525 | 29,915 | 5,000 | 29,915 |
| **All three countries** |   |   |   |   |
| Intercept | 454.591\*\*\* | 448.843\*\*\* | 415.092\*\*\* | 448.843\*\*\* |
|   | (2.045) | (3.891) | (3.159) | (6.440) |
| Class Size | 6.253\*\*\* | 8.614\*\*\* | 11.932\*\*\* | 8.614\*\*\* |
|   | (0.493) | (1.129) | (0.872) | (2.043) |
| Female | -6.307\*\*\* | -10.309\*\*\* | -7.357\*\*\* | -10.309\*\*\* |
|   | (0.988) | (1.744) | (1.385) | (0.744) |
| School Size | 0.033\*\*\* | 0.031\*\*\* | 0.040\*\*\* | 0.031\*\*\* |
|   | (0.001) | (0.002) | (0.002) | (0.003) |
| Percent female | 10.204\*\*\* | -0.087 | 26.924\*\*\* | -0.087 |
|   | (2.943) | (4.881) | (4.426) | (7.536) |
| R-Squared | 0.034 | 0.036 | 0.064 | 0.036 |
| N | 38166 | 856,554 | 15,000 | 856,554 |
| *Note.* Standard errors in parentheses. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1 |

**Table 4.** TIMSS 2015 (Grade 8) Weighted and Unweighted Single-Level (Student Level) Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Unweighted | Student Weights | Senate Weights | House Weights | Student weights +JRRWeights |
| **Canada** |   |   |   |   |   |
| Intercept | 709.763\*\*\* | 738.060\*\*\* | 738.060\*\*\* | 738.060\*\*\* | 738.060\*\*\* |
|   | (25.659) | (32.966) | (32.966) | (32.966) | (2.428) |
| Age | -8.217\*\*\* | -12.328\*\*\* | -12.328\*\*\* | -12.328\*\*\* | -12.328\* |
|   | (1.642) | (2.048) | (2.048) | (2.048) | (1.115) |
| Female | -11.678\*\*\* | -11.631\*\*\* | -11.631\*\*\* | -11.631\*\*\* | -11.631\*\* |
|   | (1.566) | (1.946) | (1.946) | (1.946) | (0.267) |
| Instruction hours per year | -0.084\*\*\* | -0.058\*\*\* | -0.058\*\*\* | -0.058\*\*\* | -0.058 |
|   | (0.010) | (0.013) | (0.013) | (0.013) | (0.015) |
| Percent female | 27.056\*\*\* | 26.142\*\* | 26.142\*\* | 26.142\*\* | 26.142\*\* |
|   | (7.946) | (10.963) | (10.963) | (10.963) | (0.456) |
| R-Squared | 0.017 | 0.016 | 0.016 | 0.016 | 0.016 |
| N | 8757 | 229,175 | 500 | 8757 | 229,175 |
| **Italy** |   |   |   |   |   |
| Intercept | 802.167\*\*\* | 812.944\*\*\* | 812.945\*\*\* | 812.944\*\*\* | 812.944\*\*\* |
|   | (35.008) | (43.954) | (43.954) | (43.954) | (6.346) |
| Age | -22.491\*\*\* | -22.568\*\*\* | -22.568\*\*\* | -22.568\*\*\* | -22.568\*\*\* |
|   | (2.333) | (2.994) | (2.994) | (2.994) | (0.219) |
| Female | -12.693\*\*\* | -11.956\*\*\* | -11.956\*\*\* | -11.956\*\*\* | -11.956\* |
|   | (2.278) | (2.656) | (2.656) | (2.656) | (1.168) |
| Instruction hours per year | 0.020\*\* | 0.015 | 0.015 | 0.015 | 0.015 |
|   | (0.010) | (0.010) | (0.010) | (0.010) | (0.004) |
| Percent female | 5.440 | -12.111 | -12.111 | -12.111 | -12.111 |
|   | (16.854) | (20.128) | (20.128) | (20.128) | (6.344) |
| R-Squared | 0.026 | 0.026 | 0.026 | 0.026 | 0.026 |
| N | 4481 | 528,839 | 500 | 4481 | 528,839 |
| **Lithuania** |   |   |   |   |   |
| Intercept | 417.471\*\*\* | 535.777\*\*\* | 535.778\*\*\* | 535.777\*\*\* | 535.777\* |
|   | (45.705) | (60.763) | (60.763) | (60.763) | (44.425) |
| Age | 3.120 | -4.837 | -4.837 | -4.837 | -4.837 |
|   | (2.969) | (3.989) | (3.989) | (3.989) | (2.501) |
| Female | -4.296\* | -6.366\*\* | -6.366\*\* | -6.366\*\* | -6.366 |
|   | (2.522) | (2.930) | (2.930) | (2.930) | (3.474) |
| Instruction hours per year | 0.031\*\*\* | 0.060\*\*\* | 0.060\*\*\* | 0.060\*\*\* | 0.060\* |
|   | (0.011) | (0.012) | (0.012) | (0.012) | (0.007) |
| Percent female | 25.989 | 0.056 | 0.055 | 0.056 | 0.056 |
|   | (16.897) | (19.328) | (19.328) | (19.328) | (3.829) |
| R-Squared | 0.003 | 0.010 | 0.010 | 0.010 | 0.010 |
| N | 4347 | 27,263 | 500 | 4347 | 27,263 |
| **All three countries** |   |   |   |   |   |
| Intercept | 707.751\*\*\* | 749.604\*\*\* | 630.004\*\*\* | 660.614\*\*\* | 749.604\*\* |
|   | (17.613) | (32.296) | (23.784) | (22.229) | (25.400) |
| Age | -13.207\*\*\* | -15.493\*\*\* | -7.444\*\*\* | -9.200\*\*\* | -15.493\* |
|   | (1.077) | (2.116) | (1.459) | (1.362) | (1.560) |
| Female | -10.001\*\*\* | -10.966\*\*\* | -9.411\*\*\* | -10.006\*\*\* | -10.966\* |
|   | (1.164) | (1.902) | (1.486) | (1.394) | (1.313) |
| Instruction hours per year | -0.014\*\*\* | -0.021\*\*\* | -0.011\* | -0.019\*\*\* | -0.021 |
|   | (0.005) | (0.008) | (0.006) | (0.006) | (0.003) |
| Percent female | 22.020\*\*\* | 0.224 | 5.072 | 10.840 | 0.224 |
|   | (6.771) | (12.433) | (9.426) | (8.744) | (0.192) |
| R-Squared | 0.012 | 0.016 | 0.006 | 0.008 | 0.016 |
| N | 17585 | 785,277 | 1,500 | 17585 | 785,277 |
| *Note.* Standard errors in parentheses. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1 |   |   |

**Table 5.** PISA 2015 Weighted and Unweighted Two-Level Random Intercept Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Unweighted | Student Weights | School Weights | Student and School Weights (Unscaled) | Student and School Weights (Size scale) | Student and School Weights (Effective scale) |
| **Canada** |   |   |   |   |   |   |
| Intercept | 477.745\*\*\* | 476.924\*\*\* | 502.557\*\*\* | 503.953\*\*\* | 502.754\*\*\* | 502.790\*\*\* |
|   | (7.645) | (9.745) | (20.361) | (20.688) | (20.354) | (20.360) |
| Class Size | 3.160\* | 3.199\* | 2.539 | 2.553 | 2.465 | 2.454 |
|   | (1.672) | (1.801) | (3.258) | (3.490) | (3.264) | (3.262) |
| Female | -2.215\* | -3.151\* | -1.878 | -2.450 | -1.771 | -1.771 |
|   | (1.219) | (1.623) | (2.318) | (1.777) | (2.320) | (2.329) |
| School Size | 0.025\*\*\* | 0.025\*\*\* | 0.016\*\* | 0.016\* | 0.016\*\* | 0.016\*\* |
|   | (0.004) | (0.004) | (0.008) | (0.009) | (0.008) | (0.008) |
| Percent female | 17.675 | 19.857 | -3.648 | -6.305 | -3.973 | -3.999 |
|   | (11.968) | (16.066) | (32.230) | (33.068) | (32.194) | (32.202) |
| ICC | 0.142 | 0.163 | 0.195 | 0.222 | 0.195 | 0.195 |
| N (School) | 759 | 759 | 759 | 759 | 759 | 759 |
| N (Student) | 20058 | 20058 | 20058 | 20058 | 20058 | 20058 |
|   |   |   |   |   |   |   |
| **Italy** |   |   |   |   |   |   |
| Intercept | 470.544\*\*\* | 468.510\*\*\* | 410.213\*\*\* | 409.177\*\*\* | 410.875\*\*\* | 410.887\*\*\* |
|   | (10.278) | (9.658) | (16.002) | (16.150) | (16.016) | (16.022) |
| Class Size | -2.327 | -2.237 | 4.699 | 4.204 | 4.292 | 4.289 |
|   | (2.641) | (2.367) | (4.788) | (4.784) | (4.789) | (4.788) |
| Female | -17.573\*\*\* | -21.122\*\*\* | -14.331\*\*\* | -18.149\*\*\* | -14.465\*\*\* | -14.492\*\*\* |
|   | (1.471) | (1.748) | (2.704) | (2.152) | (2.680) | (2.682) |
| School Size | 0.029\*\*\* | 0.029\*\*\* | 0.035\*\* | 0.035\*\* | 0.034\*\* | 0.034\*\* |
|   | (0.008) | (0.008) | (0.016) | (0.016) | (0.016) | (0.016) |
| Percent female | 11.870 | 17.331 | 28.035 | 35.050\* | 29.702 | 29.740 |
|   | (11.128) | (10.985) | (21.090) | (21.233) | (21.030) | (21.033) |
| ICC | 0.433 | 0.46 | 0.465 | 0.49 | 0.463 | 0.46 |
| N (School) | 474 | 474 | 474 | 474 | 474 | 474 |
| N (Student) | 11583 | 11583 | 11583 | 11583 | 11583 | 11583 |
|   |   |   |   |   |   |   |
| **Lithuania** |   |   |   |   |   |   |
| Intercept | 356.614\*\*\* | 357.414\*\*\* | 380.667\*\*\* | 381.016\*\*\* | 380.812\*\*\* | 380.807\*\*\* |
|   | (9.668) | (10.011) | (12.466) | (12.470) | (12.449) | (12.449) |
| Class Size | 15.380\*\*\* | 15.571\*\*\* | 9.388\*\* | 9.433\*\* | 9.457\*\* | 9.458\*\* |
|   | (2.686) | (3.142) | (4.300) | (4.267) | (4.309) | (4.310) |
| Female | -3.314\* | -7.208\*\*\* | -0.561 | -4.029\* | -0.584 | -0.602 |
|   | (1.891) | (2.203) | (2.510) | (2.136) | (2.529) | (2.528) |
| School Size | 0.047\*\*\* | 0.046\*\*\* | 0.035\*\* | 0.034\*\* | 0.034\* | 0.034\* |
|   | (0.012) | (0.012) | (0.017) | (0.017) | (0.017) | (0.017) |
| Percent female | 84.880\*\*\* | 87.463\*\*\* | 68.269\*\*\* | 71.733\*\*\* | 68.267\*\*\* | 68.279\*\*\* |
|   | (17.428) | (18.174) | (21.761) | (21.826) | (21.715) | (21.719) |
| ICC | 0.219 | 0.246 | 0.241 | 0.261 | 0.241 | 0.241 |
| N (School) | 311 | 311 | 311 | 311 | 311 | 311 |
| N (Student) | 6525 | 6525 | 6525 | 6525 | 6525 | 6525 |
|   |   |   |   |   |   |   |
| **All three countries** |   |   |   |   |   |   |
| Intercept | 441.934\*\*\* | 441.379\*\*\* | 438.160\*\*\* | 437.930\*\*\* | 438.238\*\*\* | 438.264\*\*\* |
|   | (5.210) | (6.111) | (13.923) | (14.136) | (13.893) | (13.895) |
| Class Size | 7.258\*\*\* | 7.328\*\*\* | 7.566\*\* | 7.253\*\* | 7.348\*\* | 7.340\*\* |
|   | (1.351) | (1.677) | (3.619) | (3.621) | (3.601) | (3.599) |
| Female | -6.307\*\*\* | -7.993\*\*\* | -8.558\*\*\* | -10.713\*\*\* | -8.556\*\*\* | -8.560\*\*\* |
|   | (0.849) | (1.185) | (1.733) | (1.405) | (1.725) | (1.726) |
| School Size | 0.034\*\*\* | 0.033\*\*\* | 0.014 | 0.014 | 0.014 | 0.014 |
|   | (0.004) | (0.004) | (0.011) | (0.011) | (0.011) | (0.011) |
| Percent female | 19.063\*\* | 21.723\*\* | 17.837 | 22.090 | 19.111 | 19.116 |
|   | (7.454) | (8.751) | (19.662) | (19.848) | (19.578) | (19.579) |
| ICC | 0.281 | 0.296 | 0.422 | 0.439 | 0.421 | 0.421 |
| N (School) | 1,544 | 1,544 | 1,544 | 1,544 | 1,544 | 1,544 |
| N (Student) | 38166 | 38166 | 38166 | 38166 | 38166 | 38166 |
| *Note.* Standard errors in parentheses. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1  |

**Table 6.** TIMSS 2015 (Grade 8) Weighted and Unweighted Two-Level Random Intercept Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Unweighted | Student Weights | School Weights | Student and School Weights (Unscaled) | Student and School Weights (Size scale) | Student and School Weights (Effective scale) |
| **Canada** |   |   |   |   |   |   |
| Intercept | 768.254\*\*\* | 813.241\*\*\* | 625.509\*\*\* | 707.670\*\*\* | 617.453\*\*\* | 606.263\*\*\* |
|   | (36.513) | (50.873) | (63.146) | (56.712) | (62.834) | (63.776) |
| Age | -12.804\*\*\* | -16.247\*\*\* | -6.302\*\* | -12.593\*\*\* | -5.958\*\* | -5.150\* |
|   | (1.655) | (2.790) | (3.114) | (2.741) | (3.039) | (3.109) |
| Female | -12.222\*\*\* | -13.770\*\*\* | -10.231\*\*\* | -11.886\*\*\* | -10.308\*\*\* | -10.300\*\*\* |
|   | (1.417) | (2.028) | (2.026) | (1.809) | (2.044) | (2.067) |
| Instruction hours per year | -0.085\*\*\* | -0.082\*\*\* | -0.027 | -0.024 | -0.023 | -0.023 |
|   | (0.027) | (0.030) | (0.041) | (0.041) | (0.041) | (0.042) |
| Percent female | 32.654 | 31.965 | 12.289 | 17.039 | 10.442 | 10.071 |
|   | (21.089) | (25.124) | (35.275) | (33.904) | (35.337) | (35.398) |
| ICC | 0.206 | 0.231 | 0.18 | 0.195 | 0.18 | 0.179 |
| N (School) | 276 | 276 | 276 | 276 | 276 | 276 |
| N (Student) | 8757 | 8757 | 8757 | 8757 | 8757 | 8757 |
|   |   |   |   |   |   |   |
| **Italy** |   |   |   |   |   |   |
| Intercept | 864.006\*\*\* | 850.579\*\*\* | 905.367\*\*\* | 882.923\*\*\* | 906.696\*\*\* | 906.704\*\*\* |
|   | (45.287) | (52.727) | (65.167) | (55.471) | (65.047) | (65.057) |
| Age | -25.957\*\*\* | -24.892\*\*\* | -26.486\*\*\* | -24.957\*\*\* | -26.546\*\*\* | -26.547\*\*\* |
|   | (2.215) | (2.864) | (3.264) | (2.832) | (3.259) | (3.260) |
| Female | -13.847\*\*\* | -14.011\*\*\* | -12.863\*\*\* | -13.235\*\*\* | -12.889\*\*\* | -12.890\*\*\* |
|   | (2.091) | (2.221) | (3.383) | (2.625) | (3.382) | (3.383) |
| Instruction hours per year | 0.009 | 0.009 | -0.001 | -0.001 | -0.001 | -0.001 |
|   | (0.024) | (0.020) | (0.022) | (0.022) | (0.022) | (0.022) |
| Percent female | -3.709 | -5.320 | -58.411 | -56.191 | -58.918 | -58.917 |
|   | (42.334) | (44.990) | (52.830) | (52.767) | (52.795) | (52.795) |
| ICC | 0.184 | 0.21 | 0.167 | 0.20 | 0.167 | 0.17 |
| N (School) | 161 | 161 | 161 | 161 | 161 | 161 |
| N (Student) | 4481 | 4481 | 4481 | 4481 | 4481 | 4481 |
|   |   |   |   |   |   |   |
| **Lithuania** |   |   |   |   |   |   |
| Intercept | 624.278\*\*\* | 587.883\*\*\* | 637.883\*\*\* | 605.464\*\*\* | 640.010\*\*\* | 640.177\*\*\* |
|   | (52.078) | (65.915) | (72.117) | (77.424) | (72.030) | (72.084) |
| Age | -8.117\*\*\* | -5.661 | -8.961\*\* | -6.845 | -9.101\*\* | -9.111\*\* |
|   | (2.780) | (4.044) | (4.010) | (4.437) | (4.006) | (4.010) |
| Female | -6.244\*\*\* | -6.372\*\* | -7.686\*\* | -7.137\*\* | -7.626\*\* | -7.647\*\* |
|   | (2.245) | (3.160) | (3.265) | (3.167) | (3.264) | (3.266) |
| Instruction hours per year | 0.004 | 0.005 | 0.016 | 0.017 | 0.016 | 0.016 |
|   | (0.029) | (0.025) | (0.034) | (0.033) | (0.034) | (0.034) |
| Percent female | -21.548 | -21.191 | -36.694 | -35.754 | -36.580 | -36.582 |
|   | (42.188) | (47.094) | (59.431) | (58.661) | (59.435) | (59.437) |
| ICC | 0.262 | 0.277 | 0.241 | 0.248 | 0.24 | 0.24 |
| N (School) | 208 | 208 | 208 | 208 | 208 | 208 |
| N (Student) | 4347 | 4347 | 4347 | 4347 | 4347 | 4347 |
|   |   |   |   |   |   |   |
| **All three countries** |   |   |   |   |   |   |
| Intercept | 754.296\*\*\* | 782.604\*\*\* | 813.360\*\*\* | 842.077\*\*\* | 811.009\*\*\* | 809.891\*\*\* |
|   | (24.056) | (32.246) | (44.106) | (39.159) | (43.981) | (44.327) |
| Age | -16.434\*\*\* | -18.167\*\*\* | -19.438\*\*\* | -21.529\*\*\* | -19.311\*\*\* | -19.235\*\*\* |
|   | (1.177) | (1.807) | (2.423) | (2.212) | (2.412) | (2.439) |
| Female | -11.163\*\*\* | -12.603\*\*\* | -11.318\*\*\* | -12.509\*\*\* | -11.353\*\*\* | -11.351\*\*\* |
|   | (1.044) | (1.359) | (1.999) | (1.846) | (2.000) | (2.011) |
| Instruction hours per year | -0.015 | -0.018 | -0.020 | -0.021 | -0.019 | -0.019 |
|   | (0.013) | (0.012) | (0.015) | (0.015) | (0.015) | (0.015) |
| Percent female | 11.801 | 11.472 | -22.656 | -21.030 | -23.498 | -23.571 |
|   | (18.466) | (21.380) | (30.695) | (30.651) | (30.659) | (30.673) |
| ICC | 0.235 | 0.257 | 0.184 | 0.205 | 0.183 | 0.182 |
| N (School) | 645 | 645 | 645 | 645 | 645 | 645 |
| N (Student) | 17585 | 17585 | 17585 | 17585 | 17585 | 17585 |
| *Note.* Standard errors in parentheses. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1  |

approaches were not significant. In Lithuania, for age, the estimates of school weights, size scaling, and effective scaling were significant at 0.05 level, but the unscaled approach generated a non-significant result.

Discussion

 The purpose of using sampling weights in statistical analyses in LSAS is to control for the relative contribution of different groups that make up the sample when unequal probability selection is used at different sampling stages. Since the influence of sample designs particularly the unequal probability selection on each statistical analysis is unknown, sampling weights need to be used in statistical analysis using LSAS (Rutkowski et al., 2010). However, as empirical evidence is scarce and practical guidance is insistent, it is unclear how researchers should deal with complex sampling designs and use appropriate weights variables in different statistical analyses in the context of LSAS data. This study fills this important literature gap by examining the application of various sample weights variables in conducting descriptive statistics, single-level models, and multilevel models using data from three countries and the combined data in PISA and TIMSS.

 This study has three main findings. First, in both PISA and TIMSS data, school level sampling selection and sampling weights were more important than student level selection and sampling weights. Therefore, the unweighted analyses and weighted analyses using student level specific weights generated more similar results than other weighted approaches that all include school weights. Especially in TIMSS, as only one or two intact classrooms per grade was selected on average and then all the students in the selected classroom were included, school sampling weights may have much influence on model analyses. Therefore, the TIMSS-based sampling design produced more consistent results than the PISA-based design.

 Second, with respect to descriptive statistics, weighted estimates are different from unweighted ones. However, if the descriptive statistics was conducted within each country, using students level weights such as total student weights, house weights, and senate weights or students weights plus repeated replicate weights produced same results although different weights represent sample sizes, and scaled or unscaled population sizes. However, for multi-country analysis, different sampling weights generated different estimates of means and SDs because it is the weighted estimates by the sample size in each country.

 Third, with respect to inferential statistics of model analyses, for single level model analyses, the estimates of regression coefficients were identical using different weights, but the estimates of weighted standard errors were different from unweighted estimates. Overall, the statistical significance was not that consistent across weighted and unweighted approaches for different variables of interest in different countries. For multilevel models, in both PISA and TIMSS data, the size scaling and effective scaling performed similarly. This finding is in congruence with one prior studies based on Germany data in PISA 2015 (Mang et al., 2021). In addition, this study found that the two scaling methods and the approach of using only school weights had similar results, which were different from the weighed unscaled approach. This finding is also in congruence with Mang et al. (2021). One possible reason is that student-level specific sampling weights are less important than school weights, so the difference in using different scaling methods is trivial if school-level weights were included. However, if student-level weights were used with school weights in multilevel models, scaling is essential.

 This last section provides three practical recommendations regarding the incorporation of complex sampling weights and replicate weights variables in statistical analyses using LSAS data. First, researchers need to be familiar with sampling designs and weights variables in each LSAS data. For example, it is necessary to differentiate between sampling weights and replicate weights. The latter are generated from resampling variance estimation techniques such as JRR and BRR. For example, the TIMSS 2011 data includes 80 replicate weights (W1C1-W1C80). They serve to adjust variance estimation under the influence of complex sampling designs in statistical inference. Replicate weights can be used in single-level modeling with student sampling weights, but they cannot be used with sampling weights in multilevel models.

 Second, researchers need to follow the recommendation in LSAS data user manuals that sampling weights should be used for all statistical analyses even if empirical studies showed no difference in statistical significance. Researchers may also conduct unweighted analysis as supplementary materials if needed. For descriptive statistics, for cross-country analyses, house weights or senate weights should be used to provide weighted descriptive statistics for adjusting different sample sizes in different countries. For single-country data, student weights should be used. To apply sampling weights in model analyses, if the analysis unit is at school level, school weights should be used. If analysis unit is at student level, student weights should be used. Like descriptive statistics, if researchers conduct single-level analysis within a country, student weights should be used. If the analysis involves multiple countries, house weights or senate weights can be used to appropriately adjust the influence of different sample sizes of different countries.

 Third, given that LSAS are characterized by unequal probability selection at the school level, it is essential to include school sampling weights in multilevel modeling when using LSAS data. For multilevel sampling, in the case of two-level model, the recommendation is to use student-level specific sampling weights and school weights. For the former, school weights need to be purged from student weights to get non-overlapped student-level specific weights (Shen & Konstantopoulos, 2022). In addition, either size or effective scaling for lower-level sampling weights should be used, which are provided in STATA and Mplus. If researchers are not familiar with scaling procedures or if their software program do not provide these options, they may use school sampling weights only, but it is not recommended to use student-level specific weights only or unscaled two-level weights. One caveat is that this study only used four covariates from PISA and TIMSS data in three countries. Future research may use other LSAS data from other countries or utilize more variables to verify the similarities or differences of simulation findings.

References

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods*, *35*(3), 439–460. <https://doi.org/10.1080/03610920500476598>

Binder, D. A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review / Revue Internationale de Statistique*, *51*(3), 279–292. JSTOR. <https://doi.org/10.2307/1402588>

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, *9*, 1–13.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*(1), 43–56. <https://doi.org/10.1093/biomet/73.1.43>

Koziol, N. A., Bovaird, J. A., & Suarez, S. (2017). A Comparison of Population-Averaged and Cluster-Specific Approaches in the Context of Unequal Probabilities of Selection. *Multivariate Behavioral Research*, *52*(3), 325–349. <https://doi.org/10.1080/00273171.2017.1292115>

Laukaityte, I., & Wiberg, M. (2018). Importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communications in Statistics-Theory and Methods*, *47*(20), 4991–5012. <https://doi.org/10.1080/03610926.2017.1383429>

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education*, *9*(1), 6. <https://doi.org/10.1186/s40536-021-00099-0>

OECD. (2014). *PISA 2012 technical report*. OECD Publishing.

Olson, J. F., Martin, M. O., & Mullis, I. V. (2008). *TIMSS 2007 technical report*. TIMSS & PIRLS International Study Center.

Osborne, J. (2011). Best practices in using large, complex samples: The importance of using appropriate weights and design effect compensation. *Practical Assessment, Research, and Evaluation*, *16*(1), 12. <https://doi.org/10.7275/2kyg-m659>

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(1), 23–40. <https://doi.org/10.1111/1467-9868.00106>

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 805–827. <https://doi.org/10.1111/j.1467-985X.2006.00426.x>

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117–154).

Rutkowski, L., Gonzalez, E., Joncas, M., & Von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151.

Stapleton, L. (2002). The Incorporation of Sample Weights Into Multilevel Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 475–502. <https://doi.org/10.1207/S15328007SEM0904_2>

Stapleton, L. (2013). Incorporating sampling weights into single-and multilevel analyses. In *Handbook of international large scale assessment: Background, technical issues, and methods of data analysis* (pp. 363–388).

Citation**:**

Shen, T. (2024). Statistical analyses with sampling weights in large-scale assessment and survey. *Practical Assessment, Research, & Evaluation*, 29(4). Available online: <https://doi.org/10.7275/pare.2014>

Corresponding Author:

Ting Shen

Missouri University of Science and Technology

Email: tingshen [at] mst.edu