Chapter 10
Application of Multilevel Models to International Large-Scale Student Assessment Data

Maciej Jakubowski and Tomasz Gajderowicz

Abstract This chapter discusses applications of the multilevel modeling to international large-scale student assessment (ILSA), focusing on OECD's PISA and IEA's TIMSS. Multilevel models are routinely applied to analyze these data. However, several methodological issues need to be addressed to use these models in empirical applications correctly. First, we discuss how plausible values in multilevel modeling affect estimates of fixed and random components. Second, we discuss how to consider survey weights to decompose variance and estimate separate within- and between-school effects. Third, we discuss the use of replicate weights and compare standard errors estimated with this method to those typically obtained in multilevel modeling with robust standard errors. Fourth, we discuss applications of more complex multilevel models, like three-level models and models with cross-level effects. We summarize by providing key points to consider for researchers when applying multilevel modeling with ILSA data.

Keywords International large-scale assessment · Plausible values · Survey weight · Replicate weight · Three-level model · Cross-level effect

10.1 Introduction

The educational data have a hierarchical structure as students are nested in classrooms, and classrooms are nested in schools. Thus, multilevel modeling is a natural choice for this type of data. Examples of multilevel regressions with school and student data are presented in most books discussing applications of these statistical models. Moreover, the whole structure of education systems relies on several nested layers as schools are often managed by local authorities, governed or supervised by regional or subnational entities, and finally, education systems are organized at the subnational or national level.

M. Jakubowski (✉) · T. Gajderowicz
University of Warsaw, Warszawa, Poland
e-mail: mjakubowski@uw.edu.pl

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_10

Hypotheses in empirical research in education are also often related to interactions between different governance levels. Researchers are usually interested in individual, student-level effects and between-school effects and the relationship between-country-wide policies and associations with outcomes at the local level. One can imagine adding layers related to language, culture, governance, accountability, or practices and policies. In psychometric research, models analyzing individual test or questionnaire items that are nested in students or in time periods are also applied to address issues related to measurement error.

All large-scale international assessments of students have a hierarchical structure with students nested in classrooms or schools, and then schools nested in countries. Additional layers are sometimes added when analyzing regional data, teacher effects, or item-level responses of students. Multilevel modeling with these data is popular among researchers and often involves cross-level interactions. However, important issues related to the statistical design of international large-scale assessment data need to be addressed to analyze them properly, obtain unbiased population estimates, and measure their uncertainty. This chapter discusses the usage of plausible values, survey and replicate weights, assumptions about random effects distribution, and other issues that often arise in empirical applications but are also often misunderstood or incorrectly addressed. Throughout the chapter, we provide examples using the most recent PISA and TIMSS data.

10.2 Example of Typical Use—Modeling Relationship Between Socioeconomic Background and Student Achievement in PISA 2018

There are three main advantages of multilevel modeling with large-scale student assessment data. First, they reflect the sampling structure with schools at the higher level and classrooms and students at the lower levels. We discuss below the benefits and costs of applying multilevel models to reflect the complex sampling scheme in international studies. The second advantage is that multilevel regressions provide a decomposition of the variance, and the third advantage is that they allow modeling variance at different levels, including cross-level interactions.

We use PISA 2018 data to demonstrate how to use multilevel models to analyze the relationship between student socioeconomic background and reading achievement. In education research, it is a common model used to estimate the effects net of family background. This is also a model often used in multilevel modeling textbooks, starting from the popular Raudenbush and Bryk book (2002), which opens with an example of modeling SES association with achievement. In PISA, the socioeconomic background is measured through the index of economic, social, and cultural status. This is an index that combines information about parents' education and occupation, and educational, cultural, and material resources available at home (see Avvisati,

Table 10.1 Example of multilevel analysis with PISA 2018 data—explaining reading achievement with variance decomposition and student- and school-level slopes of socioeconomic background

	(1)	(2)	(3)	(4)
	b/se	b/se	b/se	b/se
ESCS slope		16.8		
		0.66		
Between-school ESCS slope			55.1	51.6
			2.42	2.81
Within-school ESCS slope			12.5	12.5
			0.70	0.70
Constant	446.0	457.5	481.9	481.1
	2.75	2.31	1.76	2.73
Country fixed effects	No	No	No	Yes
School-level variance	5578.5	3874.4	2225.9	1904.2
	255.56	199.04	125.04	116.30
Student-level variance	6280.2	6180.8	6176.3	6175.3
	80.24	74.87	75.73	75.70
Intraclass correlation	0.47	0.39	0.26	0.24
N of schools	10,180	10,180	10,180	10,180
N of students	249,334	249,334	249,334	249,334

Note Authors' estimation with PISA 2018 microdata. Results were obtained with ten plausible values of reading achievement and a sample of all students and countries that have participated in this assessment. Student and school weights were applied with student-level weights scaled to sum to the sample size of their school

2020 for a detailed discussion of this index and its comparability across countries and Pokropek et al., 2015 for a decomposition using structural equation modeling).

Tables 10.1 and 10.2 show results for a multilevel model applied to PISA 2018 data explaining reading achievement with the ESCS index. The index was centered at the weighted school means to decompose the effects into within- and between-school effects. The estimates are based on a model with ten plausible values and different specifications regarding random components and country effects.

Table 10.1 shows results for a model with school-level random intercepts. Column (1) shows results for the empty model—the model with intercepts only. This model provides a baseline decomposition of the variance into school and student levels. One could argue that this is one of the most important findings from large-scale international assessments, that a substantial part of the total achievement variance is associated with school-level effects. In this case, for the pooled sample of all countries that participated in PISA 2018, interclass correlation shows that nearly half of the total variance is associated with schools. These estimates are obtained for a weighted sample of students with ESCS data available, so they can be used for comparisons

Table 10.2 Example of the random coefficient multilevel model with PISA 2018 data

	(1)	(2)
	b/se	b/se
Between-school ESCS slope	53.9***	55.1***
	2.30	2.43
Within-school ESCS slope	12.3***	15.3***
	0.74	0.85
Interaction between the school-average ESCS and within-school ESCS slope		5.2***
		0.96
Constant	481.1***	481.9***
	1.76	1.77
ESCS slope variance	149.2***	122.3***
	20.96	20.10
School-level variance	2250.5***	2243.2***
	125.74	126.68
Correlation (escs,constant)	0.3***	0.3***
	0.05	0.06
Student-level variance	6070.2***	6072.0***
	75.14	75.50
N of schools	10,180	10,180
N of students	249,334	249,334

Note Authors' estimation with PISA 2018 microdata. Results were obtained with ten plausible values of reading achievement and a sample of all students and countries that have participated in this assessment. Student and school weights were applied with student-level weights scaled to sum to the sample size of their school

with models incorporating ESCS effects (missing observations constituted less than 4% of the original sample).

Columns (2) and (3) compare results for an approach typical for traditional linear regression modeling with the one possible with multilevel models that decompose associations between school- and student-level effects. Column (2) shows results for a multilevel model with the ESCS index as the only explanatory variable. The slope is around 17 points meaning that one standard deviation change in the ESCS index is associated with 17 points improvement in reading scores. Column (3) shows separate estimates for between- and within-school associations of ESCS. The within-school association is slightly weaker, but the between-school association is much stronger, showing that one standard deviation increase in school-average ESCS index is associated with an improvement of more than 50 points, which is equivalent to half a standard deviation of the reading achievement distribution for OECD countries (weighting countries equally). Note also that including the school-average slope of

ESCS explains 60% of the school-level variance, while at the student level, the within-school effect explains less than 2% of the variance.

In other words, this simple model shows that socioeconomic background is a powerful predictor of average school achievement but cannot explain much of the within-school differences. While this is not a new finding in education research, international assessment data show that this relationship holds for most schools around the world. In fact, PISA data are used to compare such associations showing in which countries school composition of student socioeconomic background is a more powerful predictor of achievement and how much of the total variance is at the school level. This is a descriptive but powerful tool for comparing inequalities related to school and socioeconomic background across countries.

The last column shows a similar model but with country fixed effects. Note that the estimated coefficients for between- and within-school associations of reading achievement with the socioeconomic background are similar, so they are not driven by between-country differences in average achievement. Note also that country fixed effects are able to explain some of the school-level variance but less the school-average socioeconomic background. Again that is an interesting finding showing that differences between schools in their composition and achievement are much larger and more important than between-country achievement differences.

This model can be further expanded by slopes of explanatory variables to randomly vary and to explain this variation with, for example, cross-level interactions. Table 10.2 shows results for a model with a random coefficient for the within-school differences in the ESCS index. The model estimates the variance of intercepts at the school and student level, but also variance in the slope of the within-school ESCS index and the covariance of the ESCS slopes and school intercepts. Results show that within-school association between student ESCS and reading achievement vary significantly across schools. We hypothesize that this variation might depend on school SES composition, and this interaction is estimated by the model presented in column (2). Indeed, the higher is the average socioeconomic background of students in a school, the stronger is the relationship between within-school ESCS and achievement. This interaction effect explains around 18% of the within-school ESCS slope variation.

10.3 Applications

The above examples demonstrate the typical use of multilevel models with international large-scale student assessment data. Variance decomposition and comparisons of between- and within-school effects are commonly applied to PISA data and are similar to the first application of multilevel modeling in education (Aitkin & Longford, 1986; Raudenbush & Bryk, 1986). The approach was partly popularized by first research using PISA data (for example, Willms, 2010) and PISA OECD reports, which routinely apply these models to decompose student- and school-level relationships (see for example results presented in OECD, 2019, but also Annex A3 with notes on the technical application of these models in PISA).

The most common approach is to study school effectiveness using multilevel models with sets of school-level and student-level predictors (for a review, see Klieme, 2013). Interestingly, these models often demonstrate that learning conditions and practices at the school level are less related to achievement than student-level opinions about the teaching process. Multilevel modeling provides a unique opportunity to study this kind of question. For school-related factors, especially for studies of socioeconomic background, PISA data often provide more detailed information. For teacher-related factors, however, TIMSS and PIRLS data might be more suitable. The sampling scheme in TIMSS and PIRLS is different from whole classrooms sampled within selected schools. This opens a possibility to collect more meaningful information about teaching as questionnaires are filled by all students of a particular teacher, separately for mathematics and science. The clear link between students and their teachers opens a possibility to model this relationship with multilevel models.

The applications of multilevel modeling with PISA data go beyond typical school effectiveness research. For example, multilevel models are applied to better understand data on student wellbeing (He et al., 2019; Jakubowski & Gajderowicz, 2020; Sznitman et al., 2011), sources of bullying (Winnaar et al., 2018; Yavuz et al., 2017), and attitudes (Lu & Bolt, 2015; Pitsia et al., 2017; Sun et al., 2012). Also, the data are often combined to provide a broader picture of student achievement and related factors (for example, see Grilli et al., 2016).

The application of multilevel modeling to international student assessment data is an obvious choice, but several technical issues need to be addressed to properly estimate population relationships of interest. As we will see below, these technical issues can be addressed with a good understanding of the role of plausible values and complex sampling in deriving conclusions from multilevel models. Many statistical packages allow taking these issues into account. More complex models, for example, three-level models, are also applied to these data—however, their raise technical issues which, as discussed below, are not straightforward to address.

10.4 Plausible Values and Multilevel Models

In publicly available datasets from large-scale student assessments like PISA, TIMSS, or PIRLS, achievement results are provided as sets of the so-called plausible values. These variables reflect not only student achievement but also the uncertainty with which it is measured for the student population. Plausible values are imputations of latent student achievement. Their correct use allows obtaining unbiased estimates of achievement in student populations, correcting for measurement error when relating to other variables in standard statistical models, and obtaining proper uncertainty measures in models with student achievement (see Wu, 2005).

For some researchers, plausible values can be seen as a technical obstacle in analyzing data like PISA or TIMSS. Analysis with plausible values requires special software, commands, or the application of formulas to calculate results by hand from statistical models with separate plausible values. Thus, researchers often try to simplify the analysis with plausible values making mistakes that invalidate their results. Below we show that using plausible is quite straightforward and that common shortcuts provide highly biased results. We also show how to use plausible values to obtain initial results faster before deciding about the final model, which is often helpful in time-consuming multilevel analysis.

First, note that a single plausible value provides an unbiased point estimate. If plausible values are drawn from distributions conditional on other variables involved in the final statistical model, then correlations with single plausible values also reflect latent correlations with other variables. For example, if the final statistical model involves student gender, plausible values should be estimated based on student gender and correlation. In this case, a simple correlation between one plausible value and student gender reflects the latent correlation between gender and student achievement. Thus, estimation with one plausible value provides unbiased point estimates also for latent correlations. However, it does not capture the effect of measurement error on the estimated variance. In other words, standard errors will be downward biased as they will not reflect measurement error.

To correctly estimate point estimates and their standard errors, one needs to run separate models with each plausible value and then take the average of estimates across these models as the point estimate and use the so-called Rubin's formula to calculate their standard errors (see Rubin, 1987). A researcher can apply Rubin's formulas herself by collecting results for each plausible value and then calculating final point estimates and standard errors. Some software packages allow to use of plausible values and calculate correct results, or there are user-written packages that can do it. It is also possible to use solutions developed for multiple imputations of missing data as the formulas are the same, and correct results can be obtained after defining each plausible as an imputed variable.

Taking an intuitive uniformed shortcut by calculating first the average of all plausible values and then running statistical models with this average as a measure of student achievement is the most common mistake done by entry-level researchers. The intuition behind this approach is that the average of plausible values is still a good achievement estimate, but in reality, such a variable suffers from an artificially lower variance. Thus, depending on a model, the final results will be biased as the overall achievement variance will be underestimated, and correlations with other indicators will be overestimated (see OECD, 2009, p. 128).

Under most circumstances, it would be more advisable to use the first plausible value if it is not possible to calculate final estimates by applying Rubin's formulas to statistical models run separately with each plausible. For example, if one is mainly interested in point estimates or, for example, creates graphic illustrations of the data, using the first plausible value will suffice. Also, when exploring the data and searching for a final model, it is also advisable to use one of the plausible values to quickly run multiple models and then do proper calculations when estimating the final model. In this case, however, one should note that the results of statistical tests will be more optimistic, so with the final model, some hypotheses might be rejected even if initial findings suggest statistically significant results.

Table 10.3 illustrates the above-mentioned issues using PISA 2000 data for Poland and two-level multilevel models with students nested in schools. Models explain student reading performance but with four differently defined variables measuring achievement. The first achievement variable is the so-called Warm estimate (weighted likelihood estimate) (Warm, 1989). The results for this variable are presented in columns (1) and (5). In columns (2) and (6), results obtained with one plausible value are presented. In columns (3) and (7), the models were estimated with the average of five plausible values as the outcome variable. The columns (4) and (8) rely on Rubin's formula to calculate point estimates and standard errors from five separate multilevel models, each run with a different plausible value.

Table 10.3 Comparisons of multilevel models estimated with different measures of student achievement (PISA 2000 data for Poland and reading achievement)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Warm estimate	1st PV	Mean PV	5 PVs	Warm estimate	1st PV	Mean PV	5 PVs
	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se
PISA index of reading enjoyment					15.7	14.9	16.2	16.2
					(1.4)	(1.2)	(1.2)	(1.8)
Males (females as a base group)					0.4	0.5	0.8	0.7
					(2.6)	(2.4)	(2.2)	(2.5)
ISCED 3B schools (3A as a base group)					−57.2	−60.2	−61.3	−61.4
					(7.2)	(7.6)	(7.4)	(7.5)
ISCED 3C schools (3A as a base group)					−158.7	−168.6	−168.3	−168.3
					(7.8)	(8.3)	(8.0)	(8.2)
Constant	464.9	463.6	462.7	462.8	536.5	539.0	538.5	538.7
	(6.7)	(7.1)	(7.0)	(7.1)	(5.5)	(5.8)	(5.6)	(5.7)
School-level variance	5547.9	6220.1	6138.0	6135.3	996.2	1170.0	1100.8	1095.1
	(358.2)	(398.6)	(392.3)	(400.2)	(72.7)	(82.2)	(76.6)	(85.1)
Student-level variance	4438.3	3646.5	3177.4	3710.0	4167.6	3442.8	2947.6	3474.7
	(52.9)	(43.4)	(37.8)	(61.2)	(50.7)	(41.9)	(35.8)	(50.2)
Intraclass correlation	0.56	0.63	0.66	0.62	0.19	0.25	0.27	0.24
N of schools	127	127	127	127	127	127	127	127
N of students	3653	3654	3654	3654	3511	3512	3512	3512

Note own calculations using PISA 2000 data for Poland. Standard errors in parentheses

That point estimates for regression coefficients are highly similar across results with different achievement measures. The main difference lies in standard errors and in the estimates of variance components. In general, the Warm likelihood estimate will overestimate student achievement variance, while the variable calculated as the average of plausible values will underestimate it. The results with just one plausible value will provide unbiased point estimates and correct achievement variance estimates, but the standard errors will be underestimated as they do not reflect the measurement error. The results calculated using Rubin's formula should be taken as a reference point for other models. Note that by using standard error estimates from these methods, the precise value of both measurement and sampling error in the data can be found.

The so-called empty models presented in columns (1) to (4) show that the student-level variance is overestimated for the Warm measure and underestimated for the mean PV measure, as expected. The results with one or five plausible values are highly similar. For the school-level variance, the results are similar, although the estimate with the Warm likelihood measure of achievement seems to be lower, and the intraclass correlation is also lower, as it is based on an overestimated variance at the individual level. The intraclass correlation will be higher for the model with achievement measured as the average of plausible values as it underestimates individual variance. Regarding the standard errors, note the estimates for the individual-level predictors: index of reading enjoyment and gender. The estimates based on only one plausible value are the lowest as they do not reflect the measurement error.

10.5 Survey Weights Adjustments for Multilevel Models

Large-scale surveys, including student assessments, rely on complex sampling (stratification, two- or more sampling stages) and non-response adjustments. In general, the probability of sampling a student in school surveys will always vary. This is because sampling schemes always start with sampling schools first and then students (or whole classrooms) within schools. In this case, students from smaller schools are more likely to be selected than those from larger schools. As school size is usually correlated with important student and school characteristics, datasets obtained through such sampling schemes require weighting to correct for differences in sampling probabilities. Further corrections are applied to student- and school-level sampling probabilities due to non-response and oversampling of some populations (like private schools or minority-group students). These corrections vary across countries, and the correct use of survey weights is crucial for cross-country comparisons.

Without sampling weights, the results of the statistical model show estimates for the sample but are not representative of the population. However, the use of survey weights in multilevel modeling is not straightforward. Several methods are available to adjust for arising biases, but their performance will vary depending on cluster sizes, sampling schemes, the statistical model applied, and might even vary for various estimates from the same model (e.g., regression coefficients vs. variance components) (see Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006). A common piece of advice is to perform robustness checks to compare different approaches empirically in order to assure that results do not vary importantly, and if they do, to consider again assumptions made behind these corrections.

The probability weights for each sampling stage are required for multilevel models, and the multilevel model should reflect the hierarchical structure of the sampling design. Let's consider the simplest case with schools sampled first and then students sampled within schools. In this case, one should know the sampling probability for each school and calculate the weight as the inverse of this probability. One should also know the sampling probability after a student's school was selected and then calculate the conditional sampling weight as the inverse of this probability. Only the final combined sampling weight is available in most surveys, which reflects the inverse probability of being sampled without specifying probabilities at each sampling stage. In this case, one can calculate the conditional probability weight by dividing unconditional probability by school probability weight.

Even if sampling probabilities were available at each stage, the scale of weights at the lowest level (student level in our examples) affects the estimation of multilevel equations, which is different from standard approaches like linear regression, where the scale of weights is unimportant. Therefore, re-scaling of survey weights is necessary, but different methods can produce varying results, and it is unknown which is best fitted for the sampling scheme considered and for the analyzed population. Below we discuss three weight re-scaling methods, which are commonly applied in multilevel modeling of survey data.

As an empirical example, we estimate a two-level model with students nested in schools using the dataset from PISA 2018 with all OECD countries. The model explains student achievement using the PISA's economic, social, and cultural status (ESCS), which is an index combining information on parents' education, occupation, and family resources at home. This index highly correlates with student achievement in all countries, which is a typical finding for educational research. Students with disadvantaged backgrounds have on average lower achievement than students from privileged families. However, countries do differ in the extent to which socioeconomic background is related to achievement, which is often interpreted as a measure of inequality. A stronger relationship with performance shows larger differences in achievement depending on the socioeconomic background when compared to countries with a weaker relationship. Moreover, with multilevel models, it is possible to separate within- and between-school associations, which again can be used as a measure of segregation within and between schools by students' socioeconomic background.

Table 10.4 compares unweighted results with results weighted by student-level weights only, school-level weights only, and weighted with both student- and school-level weight with three different adjustment methods. The first method re-scales the student-level weights to be the sum of the cluster size. The second method re-scales the student-level weights to be the sum of the "effective" cluster size. These two methods do not re-scale the school-level weights, but the third method replaces the

Table 10.4 Comparison of unweighted and weighted multilevel models with different scaling methods of student-level weights—example using PISA 2018 data for OECD countries

	(1)	(2)	(3)	(4)	(5)	(6)
	No weights	Student final weights only	School weight only, no scaling	Student and school weight, scaling to cluster size	Student and school weight, scaling to effective cluster size	Student and school weight, GK scaling method
	b/se	b/se	b/se	b/se	b/se	b/se
ESCS index	19.3	37.9	16.9	16.8	16.8	19.3
	0.2	0.7	0.7	0.7	0.7	0.7
Constant	484.5	495.0	457.7	457.5	457.5	483.9
	0.6	1.1	2.3	2.3	2.3	1.5
School-level variance	2766.7	0.0	3860.4	3874.4	3873.2	3032.7
	46.7	0.0	198.0	199.0	199.0	133.3
Student-level variance	6721.9	9059.7	6178.1	6180.8	6178.4	6791.7
	21.1	202.5	74.7	74.9	74.8	92.9
Intraclass correlation	0.29	0.00	0.38	0.39	0.39	0.31
N of schools	10,180	10,180	10,180	10,180	10,180	10,180
N of students	249,334	249,334	249,334	249,334	249,334	249,334

Note All models are estimated with ten plausible values of reading achievement; dataset includes all OECD countries that have participated in PISA 2018 and have available reading achievement data

weights at the school level with the cluster averages of the combined student-level survey weights (the product of school weight and the conditional student weight) and then sets student-level weights to 1 (for detailed formulas and estimation methods see Graubard & Korn, 1996; Rabe-Hesketh & Skrondal, 2006).

The results presented in Table 10.2 demonstrate that the weighting and scaling of student-level weights play an essential role in interpreting results from the multilevel models. Column (1) shows a model without survey weights. This model shows estimates for the sample of students from OECD countries. It has no interpretation in terms of relationships in the population of OECD students. Model in column (2) uses student weights only, for which coefficients for fixed effects of ESCS index and constant are identical to standard linear regression approach. However, this model cannot properly capture variation at the school level. The multilevel models with school weights are presented in columns (3) to (6). Using school weight only provides similar results to those obtained with student weights re-scaled with different methods. That should not be surprising for PISA-based research as sampling probabilities vary mainly by a school (depending, for example, on school size) and not within schools. On the other hand, the GK method, which simply assumes that weights within schools are equal to one, provides different results with a much lower estimate of school-level variance.

Based on this example and previous research in this area, one could conclude that a researcher should use survey weights at both levels and check if different scaling methods provide similar results or should apply school weights only (see Mang et al., 2021, for a similar analysis with analogous conclusions). Scaling to cluster size or effective cluster size should provide similar results for most circumstances. Ignoring weights or using other methods might be misleading, especially when a researcher is interested not only in fixed effects but mainly in variance decomposition and associations at different levels of cross-level effects.

The research also provides little guidance for models with more levels. It is also disputable how to apply survey weights when, for example, using a three-level model with countries added as an additional layer. In this case, a typical approach taken by OECD when analyzing PISA data is to re-scale student-level weights, to sum up to the same amount for each country. Thus, the final results could be interpreted as the OECD average, and country-level effects would explain how different policies affect this average. This is relatively straightforward to apply in a linear regression model, but as we saw in the above examples, re-scaling of student weights is not trivial in multilevel models and would affect estimates of variance components.

10.6 Estimation of Standard Errors

One of the reasons for using multilevel models when analyzing international student assessment data is that they recognize clustering of students within schools (or classrooms). The so-called robust standard errors that are adjusted for a correlation of student-level observations within education institutions are optional in some software packages for standard models like linear regression but are automatically applied in multilevel models. For many researchers, this is an advantage of multilevel approaches that they also cite as an argument for using such models with international student assessment data.

Studies like PISA or TIMSS, however, rely on complex sampling schemes and non-response adjustments. Unfortunately, detailed information on sampling schemes and survey weights adjustments is not available in the documentation for the reason of confidentiality. Participating countries often ask to hide key information in this respect from the public to make it impossible, for example, to estimate achievement for subnational entities or particular groups of students. Also, many countries' personal data protection law regulations disallow to provide detailed information on sampling when it might help identify individuals. Thus, dedicated solutions are applied in international student assessments to ensure that such requirements are met. For the same reason, variables used for complex sampling and response adjustments are not provided in the datasets, so it is impossible to correct survey weights or standard errors to reflect sampling design and non-response.

In practice, a difference between estimates of standard errors obtained from multilevel models and those obtained with a methodology developed by assessment organizers will vary by country and group of students analyzed. Thus, it is an empirical question, and estimates from linear regressions that fully follow the methodology developed by IEA or OECD experts can serve as a benchmark for multilevel models. In research that is based on simpler sampling designs, such discrepancies will usually be small. However, for complex surveys like PISA or TIMSS, they might be larger for countries with a lot of non-response, hidden stratification, or oversampling of some populations.

International large-scale assessments rely on resampling methods to estimate standard errors as these methods provide several advantages. The most important is that they can be used with many statistical models, even for which complex sampling data analytical solutions do not exist. In addition, replicate weights can incorporate confidential information about sampling and non-response without revealing any details to the public. Thus, in many surveys where privacy issues are at stake, this is a method preferred over providing sampling information in the datasets or in the documentation.

The replicate weights methods developed for educational studies mimic the sampling process by dropping individual schools (primary sampling units) in each replication. Thus, they provide standard errors that take into account sampling at the school level and clustering of student observations within schools. Multilevel models take that into account by directly modeling school-level effects. Combining two approaches makes little sense, and there is little research on this topic. In practice, however, replicate weights provide additional information in studies like PISA or TIMSS, which cannot be incorporated in the multilevel models. Thus, it is an important empirical question on how results from these models compare to those obtained with replicate weights methods. In practice, if both approaches provide different standard errors, then a researcher should analyze the sampling process and information incorporated in the survey and replicate weights more carefully. When standard errors estimated using replicate weights are larger, then caution should be taken when interpreting results from multilevel models as key information about sampling or non-response corrections might affect the results.

Table 10.5 provides a comparison of different methods for calculating standard errors for similar models. As before, the model explains student reading achievement using PISA 2018 data for all OECD countries. Columns (1) to (3) provide results for linear regression models, but with fixed effects for school, so the results can be compared with the multilevel model with random school effects. In column (1), standard errors are estimated as for simple random sampling, in column (2), standard errors are corrected for clustering at the school level (sandwich estimator), and in column (3), standard errors are estimated using the Balanced Repeated Replication method with Fay's adjustment as advised in PISA technical reports (see OECD, 2020). These estimates of standard errors can be compared to those in column (4), which are estimated through the multilevel model with student and school weights and scaling to cluster size (the same model as in column 4 of Table 10.2).

Table 10.5 Comparison of standard errors obtained via different methods in linear regression and in a multilevel model—an example using PISA 2018 data for OECD countries

	(1)	(2)	(3)	(4)
	Linear regression	Linear regression with clustered standard errors	Linear regression with BRR standard errors	Student and school weight, scaling to cluster size
	b/se	b/se	b/se	b/se
ESCS index	15.5	15.5	15.5	16.8
	0.61	0.75	0.64	0.66
Constant	490.1	490.1	490.1	457.5
	0.49	0.20	1.11	2.31
School-level variance				3874.4
				199.04
Student-level variance				6180.8
				74.87
N of students	10,180	10,180	10,180	10,180
N of schools	249,334	249,334	249,334	249,334

Note All models are estimated with ten plausible values of reading achievement; dataset includes all OECD countries that have participated in PISA 2018 and have available reading achievement data

These results suggest that standard errors estimated with multilevel models are close to those obtained with the replicate weights method. In our example, more conservative are estimates with clustered standard errors in linear regression. However, a similar exercise should be performed in empirical applications to see if multilevel models provide standard errors that are more conservative than those obtained with replication weights and use additional, hidden information on the sampling process.

10.7 Multilevel Models with Additional Layers

Educational data have multiple layers, and depending on the research context and questions, these additional layers could be analyzed with multilevel models. The traditional choice of two-level models with students nested in classrooms or students nested in schools might not be optimal given that research questions might be related to other levels or interactions between these levels. For example, in research on education policy that uses international large-scale student assessment data, research questions often involve policies that are decided at the country level (e.g., the possibility of grade repetition or selection of students to different educational programs) but are applied at the school level, depending on individual teacher decisions, and are affecting relations between student-level variables and their achievement. For such research questions, it is natural to look at multilevel modeling as a perfect way to model these relationships. However, as we already saw, taking into account complex sampling is not straightforward even with two-level models. As countries vary in size and the number of schools sampled and in the population, a simple application ignoring survey weights could result in highly biased estimates.

It is questionable if country-level or any other level with a finite number of units could be modeled as a random effect. One could argue that countries or regions in which schools are nested represent observations from a superpopulation of all possible countries or regions (or policies possible to apply at these levels and randomly varying contexts). With obvious limitations of this approach, applications with more than two levels, including country or regional data, are interesting as they provide estimates of variance decomposition at these levels. For example, Grilli et al. (2016) estimate a four-level model to decompose achievement variance of 4th-grade students in Italy into the student, classroom, school, and province levels. The results show that achievement varies mostly at the individual level, and the province level is associated only with 5% or less of the overall variance. On the other hand, Hippe et al. (2018), using PISA data, show that achievement differences at the regional level in Spain and Italy are larger than differences in average achievement between EU countries. In a related paper, Hippe et al. (forthcoming) show that across the EU countries, the variance at the regional level is substantial when compared to the variance at the country level and that regional level predictors are strongly associated with regional level student achievement.

10.8 Summary

The data collected in large-scale international assessments are hierarchical in nature. The sampling scheme of these studies follows a general pattern of schools selected as primary sampling units followed by classrooms and students. Typical multilevel models applied to these data follow this sampling scheme with students nested in schools or classrooms. In this chapter, we discuss how to apply two-level models correctly with plausible values, survey weights at the school or classroom (or teacher) level, and scaling of survey weights at the student level. We also discuss issues related to standard error estimation when crucial information on sampling and non-response is hidden in replicate weights and not available for modeling in multilevel applications. Finally, we briefly discuss challenges in applying three-level models. In general, the methodology outlined in this chapter can be easily applied in popular statistical packages to properly analyze large-scale assessment data with two-level models. However, applying more complex multilevel models to these data still poses numerous challenges and requires caution when interpreting results.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (general), 149*(1), 1–26.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods, 35*(3), 439–460.

Avvisati, F. (2020). The measure of socioeconomic status in PISA: A review and some suggested improvements. *Large-Scale Assess Educ, 8*, 8. https://doi.org/10.1186/s40536-020-00086-x

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology, 9*, 49. https://doi.org/10.1186/1471-2288-9-49

Graubard, B. I., & Korn, E. L. (1996). Modeling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research, 5*, 263–281. https://doi.org/10.1177/096228029600500304

Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2016). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modeling of student achievement. *The Annals of Applied Statistics, 10*(4), 2405–2426.

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy &amp; Practice, 26*(4), 369–385.

Hippe, R., Jakubowski, M., & De Sousa Lobo Borges De Araujo, L. (2018). *Regional inequalities in PISA: The case of Italy and Spain*, EUR 28868 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-79-76296-3. https://doi.org/10.2760/495702, JRC109057.

Hippe, R., Jakubowski, M., & De Sousa Lobo Borges De Araujo, L. (forthcoming). *Regional variation of student performance in Europe: A multilevel model using unique PISA regional data.*

Jakubowski, M., & Gajderowicz, T. (2020). Student well-being factors: A multilevel analysis of PISA 2015 international data. *European Research Studies Journal, 23*(4), 1312–1333.

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115–147). Springer.

Lu, Y., & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-Scale Assessments in Education, 3*(1), 1–18.

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modeling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education, 9*(1), 1–39.

OECD. (2009). *PISA data analysis manual: SPSS* (2nd ed.). PISA, OECD Publishing, Paris.

OECD. (2019). *PISA 2018 results (volume III): What school life means for students' lives*, PISA, OECD Publishing. https://doi.org/10.1787/acd78851-en

OECD. (2020). *PISA 2018 technical report*. OECD Publishing, Paris. Available at https://www.oecd.org/pisa/data/pisa2018technicalreport/

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, 60*, 23–40. https://doi.org/10.1111/1467-9868.00106

Pitsia, V., Biggart, A., & Karakolidis, A. (2017). The role of students' self-beliefs, motivation and attitudes in predicting mathematics achievement: A multilevel analysis of the Programme for International Student Assessment data. *Learning and Individual Differences, 55*, 163–173.

Pokropek, A., Borgonovi, F., & Jakubowski, M. (2015). Socioeconomic disparities in academic achievement: A comparative analysis of mechanisms and pathways. *Learning and Individual Differences, 42*, 10–18. https://doi.org/10.1016/j.lindif.2015.07.011

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*, 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* (2nd ed.). Sage.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 1–17.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons Inc., New York.

Sun, L., Bradley, K. D., & Akers, K. (2012). A multilevel modeling approach to investigating factors impacting science achievement for secondary school students: PISA Hong Kong sample. *International Journal of Science Education, 34*(14), 2107–2125.

Sznitman, S. R., Reisel, L., & Romer, D. (2011). The neglected role of adolescent emotional wellbeing in national educational achievement: Bridging the gap between education and mental health policies. *Journal of Adolescent Health, 48*(2), 135–142.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record, 112*, 1008–1037.

Winnaar, L., Arends, F., & Beku, U. (2018). Reducing bullying in schools by focusing on school climate and school socioeconomic status. *South African Journal of Education, 38*(1).

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Yavuz, H. Ç., Demirtasli, R. N., Yalcin, S., & Dibek, M. İ. (2017). The effects of student and teacher level variables on TIMSS 2007 and 2011 mathematics achievement of Turkish students. *Egitim ve Bilim, 42*(189).

Maciej Jakubowski is a researcher in education and labor market policy and policy advisor. He holds a Ph.D. degree in economics and an M.A. in sociology from the University of Warsaw, where he works as an assistant professor. Between 2008 and 2012, he has worked for the PISA team at the OECD. Since 2012 served as an under-secretary of state at the Polish Ministry of Education. In 2014, he established Evidence Institute to promote evidence-based practice and support countries in analyzing international student assessments. His academic research focuses on large-scale student assessments and the methodology of policy evaluation.

Tomasz Gajderowicz is a researcher and policy advisor in the field of education and the labor market. Tomasz specializes in microeconometric methods for measuring incentives and preferences. Tomasz holds a Ph.D. in economic sciences and works as a consultant for the European Commission, World Bank, and other national and international institutions. He works as an assistant professor at the University of Warsaw and serves as a Research Director at Evidence Institute Foundation. He is the author of several publications about the transition from education to the labor market and research methodology.