Journal of Statistical Software

July 2025, Volume 113, Issue 5. doi: 10.18637/jss.v113.i05

SURVEYHLM: A SAS Macro for Multilevel Analysis with Large-Scale Educational Assessment Data

Daniel Kasper
UHH Universität Hamburg

Katrin Schulz-Heidorf
UHH Universität Hamburg

Knut Schwippert
UHH Universität Hamburg

Abstract

Special techniques must be considered during analysis of large-scale educational assessment (LSA) data. In this regard, many software packages are available to support researchers conducting secondary analyses. However, the software packages available for multilevel analyses are somewhat limited and usually contain only a few of the required techniques. In this article, we review the technical details of LSA studies and describe our comparison of software for multilevel analyses by questioning the extent to which these packages take these technical details into account. In accordance with our findings from this comparison, we developed a SAS macro for multilevel analyses of LSA data that meets all technical requirements. The macro **SURVEYHLM** fits multilevel models with LSA datasets. **SURVEYHLM** can handle up to three levels. It can fit different correlation structures for the random components and use plausible values as response variables, and the responses do not necessarily need to be normally distributed. Weights can be specified on levels 1, 2 and 3. Scaling of the level-specific weights is possible, and standard errors can be based on a sandwich estimator or calculated with either the jackknife replication technique or through user-supplied replication weights. Examples of applications are given.

Keywords: large-scale educational assessment, TIMSS, PIRLS, PISA, GLMM, multilevel analyses.

1. Introduction

International large-scale assessment studies (LSAs) of education measure students' educational achievement and provide researchers and policymakers with vital information on educational performance over time and across countries. Students are tested in domains such as reading literacy, mathematics and science literacy and, more recently, civic knowledge and computer and information literacy (Foy 2017, 2018; Jung and Carstens 2015; Köhler, Weber,

Brese, Schulz, and Carstens 2018; OECD 2009). Each program of studies usually follows a three- to five-year cycle and assesses student achievement at important stages of their education – at the end of primary and at the end of compulsory education. The International Association for the Evaluation of Educational Achievement (IEA) has been conducting large-scale international assessments of educational achievement for more than 50 years (Husén and Postlethwaite 1996). Among these assessments are the Trends in International Mathematics and Science Study (TIMSS; Mullis, Martin, Foy, and Hooper 2016a,b), which measures the mathematics and science literacy of fourth- and eighth-grade students; the Programme in International Reading Literacy Study (PIRLS; Mullis, Martin, Foy, and Hooper 2017), which assesses the reading literacy of fourth-grade students; the International Computer and Information Literacy Study (ICILS; Fraillon, Schulz, and Ainley 2013) for Grade 8 students; and the International Civic and Citizenship Study (ICCS; Schulz, Ainley, Fraillon, Losito, Agrusti, and Friedman 2018), which assesses the civic knowledge and attitudes of students typically enrolled in Grade 8. Another well-known LSA study of student achievement is the Programme for International Student Assessment (PISA; OECD 2016a,b), conducted by the Organisation for Economic Co-operation and Development (OECD). It assesses the mathematics, science, and reading literacy of 15-year-old students.

Since the first international educational comparative studies were conducted in the 1960s (Husén and Postlethwaite 1996), progress in LSAs has been accompanied by methodological advances in domains such as psychometrics, sampling theory, test development, statistics and, more recently, the use of digital technology for collecting assessment data (Masters 2017). However, these advances, as well as the specific characteristics of the respective datasets, pose challenges for researchers, especially those who are used to more traditional dataset structures. During LSA studies, simple random sampling is rarely used to select the participants (e.g., students) from the population of interest. At the same time, those conducting the assessments usually endeavor to cover a broad range of content relating to the subject and the grade level being assessed (e.g., mathematics). This aim led to the establishment of rotated test designs, where each participating student answers only a small number of the total available number of survey/questionnaire items (Berezner and Adams 2017; Rutkowski, Gonzales, Joncas, and Von Davier 2010).

These rotated test designs, also known as matrix-sampling designs (Von Davier, Sinharay, Oranje, and Beaton 2007), usually combine sets of items into blocks of equal administration time, with each set allocated to one block only. These blocks are then combined into test booklets in a rotated manner, with blocks overlapping the booklets. While this approach minimizes the test burden and subsequent consequences such as fatigue, it poses challenges for data handling (Caro and Biecek 2017). Matrix-sampling designs also require researchers to use complex statistical techniques to estimate student performance at the population level so that they can make inferences of the kind they would make if students had responded to the whole assessment (Caro and Biecek 2017; Von Davier, Gonzalez, and Mislevy 2009).

All LSA datasets, publicly available through download via the IEA and OECD websites, are accompanied by information necessary to correctly handle their technically complex structures. However, the complexity of these approaches poses challenges for researchers because they need knowledge of statistical concepts and computational methodologies that are usually beyond the scope of the typical secondary analyst (Mislevy 1991; Rutkowski *et al.* 2010). Consequently, statistical programs that handle these special features of LSA data and allow for complex models such as multilevel ones while simultaneously being easy to handle are

essential for the continuation of methodologically sound research within the context of international large-scale educational assessments. This article presents a SAS macro that fulfills these necessities of methodological accuracy and usability while enabling researchers to concurrently estimate complex models such as those developed during multilevel analyses. The code for this macro as well as the code for the calculated examples and the data sets are available on the pages of the Journal of Statistical Software.

2. Multilevel analysis with large-scale data

Sampling designs for all surveys used in this article, such as IEA's TIMSS (Mullis and Martin 2013), PIRLS (Mullis and Martin 2015), and ICILS (Fraillon *et al.* 2013), or the OECD's PISA (OECD 2017a), include stratification, clustering, and unequal selection probabilities (the technical details of these studies can be found in Fraillon, Schulz, Friedman, Ainley, and Gebhardt 2015; Martin and Mullis 2012; Martin, Mullis, and Hooper 2016; OECD 2017b). A sample that is drawn in accordance with this type of design is called a complex sample (Mislevy 1991). Statistical analysis of complex survey sample data has to take the attributes of these data into account.

Multilevel models, such as the linear mixed model (LMM),[1] are often used to analyze LSA datasets (Atar and Atar 2012; Boulifa and Kaaouachi 2015; Caponera and Losito 2016; Cosgrove and Cunningham 2011; Demir, Kiliç, and Ünal 2010; Ghagar, Othman, and Mohammadpour 2011; Grilli, Pennoni, Rampichini, and Romeo 2015; Ismail, Samsudin, and Zain 2014; Leino and Malin 2006; Liou and Hung 2015; Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, and Muthén 2008; Martin, Foy, Mullis, and O'Dwyer 2013; Meroni, Vera-Toscano, and Costa 2015; Mohammadpour 2013; Mohammadpour, Kalantarrashidi, and Shekarchizadeh 2015; OECD 2016a,b; Smith, Wendt, and Kasper 2016; Sun, Bradley, and Akers 2012; Tavsancil and Yalcin 2015; Webster and Fisher 2000; Wendt, Kasper, and Trendtel 2017; Wiberg and Rolfsman 2013). However, software that integrates all the essential statistical techniques usually applied during analyses of these datasets (e.g., plausible values, sampling weights, and replication weights; see below) is either rare, integrates only some of these features (which can result in biased estimates), or can be used only for datasets from certain LSA studies (see Section 3). These reasons explain why we developed the SAS **SURVEYHLM** macro. Based on the generalized linear mixed model (GLMM; McCulloch *et al.* 2008), the macro fits multilevel models that take the aforementioned features of LSA datasets into account in order to obtain appropriate estimates for the population of interest.

In the following sections of this article, we briefly overview the GLMM and the essential techniques that researchers usually apply when analyzing LSA datasets. Because multilevel modeling is well documented in the literature (Bickel 2007; De Leeuw and Meijer 2008; Hox 2010; Kreft and De Leeuw 1998; Raudenbush and Bryk 2002; Skrondal and Rabe-Hesketh 2004; Snijders and Bosker 2012), we do not provide background information on this process.

[1] In educational research, the statistical method is often called hierarchical linear modeling (HLM; Raudenbush and Bryk 2002). However, as Woltman, Feldstain, MacKay, and Rocchi (2012) have pointed out, the development of this method occurred simultaneously across many fields, and it is known by several names, among others, multilevel-, mixed linear-, mixed effects-, and covariance components-modeling (Hofmann 1997; Raudenbush and Bryk 2002; Woltman *et al.* 2012). We used the linear mixed model framework in this paper, not only because it can be seen as a special case of the generalized linear mixed model (GLMM; McCulloch, Searle, and Neuhaus 2008), but also because we needed it to introduce (later in this article) multilevel modeling for non-normally distributed response variables.

We do, however, show how the GLMM can be used to fit multilevel models. Also, because most of the techniques that are usually applied to LSA datasets have been well documented and discussed (see Beaton and Johnson 1992; Johnson and Rust 1992; Martin *et al.* 2016; Martin and Mullis 2012; Mislevy 1991; Mislevy, Beaton, Kaplan, and Sheehan 1992a; Mislevy, Johnson, and Muraki 1992b; OECD 2017b; Wolter 2007), our overview of these techniques is short and mainly oriented on the techniques implemented in the TIMSS, PIRLS, and PISA surveys (Martin *et al.* 2016; Martin and Mullis 2012; OECD 2017b). Furthermore, because the statistical techniques used in other LSA studies are very similar, we consider that restricting our focus to these studies is acceptable.[2] We begin our overview by introducing the proficiency estimation that uses plausible values. We then discuss the use of sample weights and the procedure for estimating the standard errors of all proficiency statistics. We end our overview with an introduction to the GLMM followed by a short description of the technical details of the **SURVEYHLM** macro.

2.1. Proficiency estimation using plausible values

As described above, LSA studies of educational achievement usually use a matrix-sampling design to assign assessment blocks to student booklets. The immediate result of the matrix-sampling design is that the raw scores for different students (i.e., the scores from the different test booklets) are not directly comparable. Instead, a linking device needs to be implemented. Large-scale educational assessment studies rely on item response theory (IRT) scaling to combine and link student responses (Martin *et al.* 2016; Martin, Mullis, and Hooper 2017; OECD 2017b; Von Davier, Carstensen, and Von Davier 2008; Von Davier *et al.* 2007). In this regard, it is assumed that the conditional probability of item response x_i to item i can be expressed as $p(x_i \mid \boldsymbol{\theta}, \boldsymbol{\beta}_i)$, with the $d \times 1$ latent parameter vector $\boldsymbol{\theta}^\top = (\theta_1 \quad \cdots \quad \theta_d)$ and the $t_i \times 1$ item parameter vector $\boldsymbol{\beta}_{iT}^\top = (\beta_{i1} \quad \cdots \quad \beta_{it_i})$.[3]

The function assumed for $p(x_i \mid \boldsymbol{\theta}, \boldsymbol{\beta}_i)$ depends on the LSA study, its cycle of assessments, and the scoring of the items. For example, TIMSS and PIRLS use a three-parameter logistic model (3PL; Birnbaum 1968) for dichotomously scored multiple-choice items, a two-parameter logistic model (2PL; Birnbaum 1968) for dichotomously scored constructed-response items, and the partial credit model (PCM; Masters 1982; Wright and Masters 1982) for polytomous constructed-response items (Martin *et al.* 2016, 2017). In contrast, PISA 2000 to 2012 cycles fitted the Rasch model (RM; Rasch 1960) to dichotomously scored items and used the PCM for items with multiple score categories (OECD 2014). Since PISA 2015, the 2PL model has been used for dichotomously scored responses and the generalized partial credit model (GPCM; Muraki 1992) for items with more than two ordered response categories (OECD 2017b).

A student's response to any subset of items induces a likelihood function for $\boldsymbol{\theta}$. However, as mentioned above, the matrix-sampling design used in LSA studies means that each student answers only a portion of all items. Although the actual number of items per scale in a booklet differs somewhat from study to study, the LSA test booklets usually have no fewer than 12 items or no more than 20 items per scale. Hence, point estimates for $\boldsymbol{\theta}$, such as the

[2] We will, however, briefly mention the differences between the studies when outlining important steps in the implemented techniques.

[3] d is the dimension of the latent proficiency. For example, PIRLS assumes that reading purpose consists of the two dimensions *literary experience* and *acquire and use information*. Consequently, in this case, $d = 2$ (Martin *et al.* 2017).

MLE $\hat{\boldsymbol{\theta}}$ or the Bayes mean estimate $\bar{\boldsymbol{\theta}} = \mathsf{E}(\boldsymbol{\theta} \mid \boldsymbol{x})$, may not be very precise. Therefore, instead of estimating individual values for $\boldsymbol{\theta}$, LSA researchers typically estimate parameters $\boldsymbol{\alpha}$ of a population distribution $p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$ of $\boldsymbol{\theta}$, and use missing value theory (Rubin 1987) to do this. According to Mislevy (1991), we can consider $\boldsymbol{\theta}$ as a variable whose responses are missing for all respondents. Also, because missingness does not depend on the value of $\boldsymbol{\theta}$, missing at random (MAR) is assumed, which means the response mechanism can be ignored (Rubin 1987). Under these conditions the expectation of a statistic $q = q(\boldsymbol{\theta}, \boldsymbol{W})$, that is, a statistic given $\boldsymbol{\theta}^\top = \begin{pmatrix} \theta_1 & \cdots & \theta_N \end{pmatrix}$ and background variables $\boldsymbol{W}$, where $\boldsymbol{W}^\top = \begin{pmatrix} \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_N \end{pmatrix}$ and $\boldsymbol{w}_j^\top = \begin{pmatrix} w_{j1} & \cdots & w_{ju} \end{pmatrix}$, can be written as

$$\mathsf{E}(q \mid \boldsymbol{X}, \boldsymbol{W}) = \int q(\boldsymbol{\theta}, \boldsymbol{W}) p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{W}) d\boldsymbol{\theta}. \tag{1}$$

LSA studies typically collect hundreds of variables $\boldsymbol{g}_j$ for each student. These variables are used to perform a principal component analysis. A number u of principal components (the $\boldsymbol{w}_j$) that explain, say, 90% of the variation in the $\boldsymbol{g}_j$ is then used for further analysis. In simple terms, and in accordance with the results of Rubin (1987), Equation 1 can be approximated by taking $m = 1, \cdots, M$ random samples from $p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{W})$, calculating $q(\boldsymbol{\theta}, \boldsymbol{W})$ for each of these sampled values, and then taking the average of these M statistics as an estimate for $\mathsf{E}(q \mid \boldsymbol{X}, \boldsymbol{W})$.

LSA studies typically factor the distribution $p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{W})$ into two components (Martin *et al.* 2016, 2017; OECD 2014, 2017b) and use these to build the posterior distribution for θ_j

$$p(\theta_j \mid \boldsymbol{x}_j, \boldsymbol{w}_j, \boldsymbol{\gamma}, \sigma^2) = \frac{p(\boldsymbol{x}_j \mid \theta_j, \boldsymbol{\beta}) p(\theta_j \mid \boldsymbol{w}_j, \boldsymbol{\gamma}, \sigma^2)}{\int p(\boldsymbol{x}_j \mid \theta_j, \boldsymbol{\beta}) p(\theta_j \mid \boldsymbol{w}_j, \boldsymbol{\gamma}, \sigma^2)}. \tag{2}$$

The first factor $p(\boldsymbol{x}_j \mid \theta_j, \boldsymbol{\beta})$, the latent variable model, corresponds to the IRT model. The second factor $p(\theta_j \mid \boldsymbol{w}_j, \boldsymbol{\gamma}, \sigma^2)$, the population model, is the prior distribution of θ_j. If $d = 1$, a normal distribution with a latent regression model is assumed for the population model, that is, $\theta_j \mid \boldsymbol{w}_j \sim N(\boldsymbol{w}_j^\top \boldsymbol{\gamma}, \sigma^2)$, and a multivariate normal distribution is assumed for multidimensional $\boldsymbol{\theta}_j$. Thus, $\boldsymbol{\theta}_j \mid \boldsymbol{w}_j \sim N(\boldsymbol{\Gamma}^\top \boldsymbol{w}_j, \boldsymbol{\Sigma})$, with $\theta_j^\top = \begin{pmatrix} \theta_{j1} & \cdots & \theta_{jd} \end{pmatrix}$. Note that the mean parameter vector $\boldsymbol{\gamma}$ (or matrix $\boldsymbol{\Gamma}$) and the variance σ^2 (or variance matrix $\boldsymbol{\Sigma}$) are assumed to be the same for all respondents. LSA studies use a three-stage estimation process for the sample from Equation 2 (Von Davier *et al.* 2007), and they repeat this three-step procedure M times, a process that results in M sets of imputations (plausible values) for each student in the sample. The values for M may differ from study to study. For example, $M = 10$ in PISA 2015 (OECD 2017b), while $M = 5$ in TIMSS 2015 (Martin *et al.* 2016).

LSA studies use the plausible values to evaluate Equation 1 for an arbitrary function q. If we assume q_m is such a statistic based on the m plausible value $\boldsymbol{\theta}_j^m$, and s_m^2 is the corresponding sample variance, then the best estimate of q obtainable from the plausible values is

$$\mathsf{E}(q \mid \boldsymbol{X}, \boldsymbol{W}) \approx \hat{q} = \frac{\sum q_m}{M},$$

with standard error

$$\begin{aligned} s_{\hat{q}}^2 &= \frac{\sum s_m^2}{M} + (1 + M^{-1}) \frac{\sum (q_m - \hat{q})^2}{M - 1}, \\ &= \bar{U} + (1 + M^{-1}) B, \end{aligned} \tag{3}$$

where $\bar{U}$ reflects the uncertainty due to sampling of students from the population, and B reflects the uncertainty because of $\boldsymbol{\theta}$ having been imputed M times.

We need to note, though, that LSA studies generally also use ordinal variables z_j^m as the dependent variables (e.g., *benchmark values* in TIMSS and PIRLS or *proficiency levels* in PISA; Martin *et al.* 2016, 2017; OECD 2017b). In simple terms, these ordinal variables are transformations of the continuous plausible values θ_j^m, in the sense that $z_j^m = c$ when $k_l^c \leq \theta_j^m < k_h^c$ (TIMSS and PIRLS) or $k_l^c < \theta_j^m \leq k_h^c$ (PISA), where the k's are ordered cut-off values on the continuous scale of θ_j^m (i.e., $k_l^c < k_l^{c+1}$ and $k_h^c < k_h^{c+1}$), and where $c = 1, 2, \ldots, C$ is, for example, the code for the benchmark value in TIMSS and PIRLS or the number of proficiency levels in PISA (Foy 2017; OECD 2017b). The z^m variables describe what students typically know and can do at a given level of θ^m. A typical research question might be that of asking how many students within a sample of students performed at a given level c, and whether that proportion was associated with gender.

2.2. Sampling weights

Although LSA studies use systematic random sampling to select students for the survey, weights must be incorporated into the analysis to compensate for these different selection probabilities. Several reasons lead to the need to vary weights across students, for example, school, class, or student nonresponse. LSA studies use different methods to determine the survey weights, with the choice of method depending on each study's sample design (see Section 1). However, the overall student sampling weight is usually a composite of level-specific weights, possibly adjusted for unequal selection probabilities and/or a trimming factor.

For example, in TIMSS 2015 the overall student sampling weight W_{ils} for student i in classroom l of school s was

$$W_{ils} = W_{i|ls} \times W_{l|s} \times W_s.$$

Here, $W_{i|ls}$ is the final student weight for students in classroom l of school s, $W_{l|s}$ is the final class weight of class l in school s, and W_s is the final school weight of school s (Martin *et al.* 2016). The final weights are basically the (nonresponse) adjusted inverse (conditional) selection probabilities of the respective unit. In PISA 2015, however, the sampling of students within schools did not take the class level into account, and the overall student sampling weight was therefore a composite of just two components (OECD 2017b). Given a student i in school s, we can write the overall student weight as

$$W_{is} = t_2 W_{i|s} \times t_1 W_s,$$

where $W_{i|s}$ is the adjusted final student weight, W_s is the adjusted final school weight, and t_2 and t_1 are trimming factors used to reduce exceptionally large weights.

Various authors favor the use of level-specific weights during multilevel analyses (Asparouhov 2006; Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash 1998; Rabe-Hesketh and Skrondal 2006; Rutkowski *et al.* 2010). These level-specific weights can be calculated by, for example, combining the final weights appropriately. Consider a three-level analysis of TIMSS 2015 data, where students are on level 1, classes are on level 2, and schools are on level 3. Here, the weight for level 1 could be $W_{i|ls}$, the weight for level 2 could be $W_{l|s}$, and the weight for level 3 could be W_s. When researchers use level-specific weights, they usually scale them

to further reduce small sample biases (Asparouhov 2006; Carle 2009; Pfeffermann *et al.* 1998; Rabe-Hesketh and Skrondal 2006).

2.3. Procedure for estimating standard errors

As can be seen in Equation 3, the standard error for estimates based on plausible values has two components. The first reflects the uncertainty due to the sampling process (i.e., sampling variance), while the second reflects the uncertainty due to the estimation of the plausible values (i.e., imputation error or measurement error). LSA studies rely on resampling schemes to estimate the sampling variance s^2. Although the different LSA studies use different resampling schemes, for example, the balanced repeated replication (BRR) technique in PISA (OECD 2017b) and the jackknife repeated replication (JRR) technique in TIMSS and PIRLS (Martin *et al.* 2016, 2017), the procedure for estimating the sampling variance can generally be described as follows:

1. Countries with a two-stage sampling design (i.e., where schools are sampled during the first stage and students in these sampled schools are sampled during the second stage) paired schools on the basis of the explicit and implicit stratification and frame ordering (e.g., measure of size) used in the sampling.[4] Countries with a single-stage sampling design (e.g., sampled only students) or a three-stage sampling design (e.g., first sampled regions, then schools, then students) did the pairings at this level and then adjusted the remaining steps (if applicable) accordingly. The literature refers to these pairs as variance strata or zones or pseudo-strata (Adams and Wu 2002; Judkins 1990; Rust 1985; Rust and Rao 1996; Wolter 2007).

2. The zones are numbered sequentially from $h = 1, \ldots, H$.[5]

3. Within each zone, one school is randomly numbered $j = 1$; in the other $j = 2$.[6]

4. The information H on school zone and the information on school number $j = 1, \ldots, J$ are attached to the data for the sampled students.

5. A set of $n_{rw} = 1, \ldots, N_{rw}$ replication weights based on the zone number and the school number are constructed, with the value for N_{rw} depending on the study and the cycle. For example, $N_{rw} = 80$ in PISA 2015 (OECD 2017b), $N_{rw} = 75$ in TIMSS 2011 (Martin and Mullis 2012), and $N_{rw} = 150$ in TIMSS 2015 (Martin *et al.* 2016). The replication weight for zone h, school j, and student i is constructed as

$$W_{hji} = c_{hj} \times W_{0i},$$

where c_{hj} is the replication factor and W_{0i} is the overall sampling weight of student i (possibly adjusted for nonresponse, given replicate n_{rw}).

[4]The procedure differs if the number of schools in an explicit stratum is an odd number. For example, PISA formed a triple of schools (OECD 2017b) while TIMSS and PIRLS randomly divided the students in the remaining school into two *quasi* schools (Martin *et al.* 2016, 2017).

[5]The value for H may differ across studies. For example, $H = 75$ in TIMSS 2015.

[6]In triplets, the third school is indexed by $j = 3$.

6. The procedures used to construct the replication factors c_{hj} also differ in accordance with the study and cycle. For example, in TIMSS 2015 (Martin *et al.* 2016) and PIRLS 2016 (Martin *et al.* 2017), the replicate factors were

$$c_{jh} = \begin{cases} 2 & \text{for students in school } j \text{ of sampling zone } h, \\ 0 & \text{for students in the other school of sampling zone } h, \\ 1 & \text{for students in any other sampling zone,} \end{cases}$$

while in TIMSS 2011 and PIRLS 2011 (Martin and Mullis 2012) they were

$$c_{jh} = \begin{cases} 2 & \text{for students in school } j = 1 \text{ of sampling zone } h, \\ 0 & \text{for students in school } j = 2 \text{ of sampling zone } h, \\ 1 & \text{for students in any other sampling zone.} \end{cases}$$

A comparison of the last two approaches shows that the difference between TIMSS 2015/PIRLS 2016 and TIMSS 2011/PIRLS 2011 is the following: Each of the studies had only two schools per sampling zone. However, two replication weights per sampling zone were constructed in later cycles of these studies, and only one replication weight per sampling zone was constructed for TIMSS 2011/PIRLS 2011. Hence, this approach has similarities to the JK2 procedure, as explained in, for example, Rust and Rao (1996). However, unlike the JK2, the designation of schools (i.e., $j = 1$ and $j = 2$) is not random but is given as a fixed indicator, as evident in the datasets of TIMSS 2011/PIRLS 2011 and the earlier cycles of these studies. In both the IEA IDB Analyzer software distributed with the TIMSS and PIRLS datasets (Foy, Arora, and Stanco 2013) and the macro **SURVEYHLM**, this procedure can be invoked with `jktyp = half`. In PISA 2015 (OECD 2017b), one of the two schools in each zone received a replicated factor of $c_{hj} = 1.5$ and the remaining schools $c_{hj} = 0.5$.[7] During this step, entries in a Hadamard matrix of order 80 are used to determine which schools receive inflated weights and which receive deflated weights.

7. A set of N_{rw} replicate estimates, $q_{n_{rw}}$, is created through use of the corresponding replication weight W_{hji}.

8. If q is the estimate of a given statistic from the full sample of students, then the estimate of the sampling variance s^2 for that statistic is given by

$$\mathsf{VAR}(q) = s^2 = k \sum_{n_{rw}=1}^{N_{rw}} \left\{ (q_{n_{rw}} - q)^2 \right\}, \tag{4}$$

where the value of k depends on the study. For example, in TIMSS 2015, $k = 0.5$ while in TIMSS 2011 $k = 1$ (Martin and Mullis 2012; Martin *et al.* 2016).

9. The square root of s^2 is used as the appropriate estimate for the standard error of any statistic derived from the variables other than plausible values.

[7] For triplets, one of the schools (designated as random) received a factor of $c_{hj} = 1.7071$ for a given replicate and the other two schools received a factor of $c_{hj} = 0.6464$. Alternatively, one school received a factor of $c_{hj} = 0.2929$, and the other two schools received a factor of $c_{hj} = 1.3536$.

10. For plausible values, either step 7 and 8 are repeated M-times (one cycle for each plausible value), yielding estimates s_m^2, or only one time via use of the first plausible value (thus yielding s_1^2). These estimates are then used as the basis for computing $\bar{U}$ (adjusting M to 1 for s_1^2), and $\bar{U}$ is used in Equation 3 to calculate $s_{\hat{q}}^2$. The square root of $s_{\hat{q}}^2$ yields the standard error for the statistic.

2.4. Generalized linear mixed model for multilevel analysis

In the following, we assume that we have $i = 1, \ldots, n$ observations on response variable y_i with $\boldsymbol{y} = (y_1 \quad \cdots \quad y_n)^\top$. The typical assumptions about y_i that are associated with generalized linear mixed models (GLMMs) are:

1. $y_i \mid \boldsymbol{\gamma} \sim$ indep. $f(y_i \mid \boldsymbol{\gamma})$,
2. $f(y_i \mid \boldsymbol{\gamma}) = h(y_i) \exp[\{\eta_i T(y_i) - A(\eta_i)\}/\phi]$,
3. $\mathsf{E}[\boldsymbol{y} \mid \boldsymbol{\gamma}] = g^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma})$, and
4. $\boldsymbol{\gamma} \sim w(\boldsymbol{\gamma})$,

with $(\partial A(\eta_i)/\partial \eta_i) = \mu_i = \mathsf{E}[y_i \mid \boldsymbol{\gamma}]$, and $(\partial^2 A(\eta_i)/\partial \eta_i^2)\phi = \sigma^2_{y_i|\boldsymbol{\gamma}}$ (Breslow and Clayton 1993; Karim and Zeger 1992; McCullagh and Nelder 1989; McCulloch *et al.* 2008; Pfeffermann *et al.* 1998; Pinheiro and Bates 1995; Rabe-Hesketh and Skrondal 2006; Tuerlinckx, Rijmen, Verbeke, and De Boeck 2006; Wolfinger and O'Connell 1993).[8] Thus, a typical assumption when the random vector $\boldsymbol{\gamma}$ is given is that the elements y_i are independent and that each element has a distribution $f(y_i \mid \boldsymbol{\gamma})$.[9] A second assumption is that a differentiable monotonic link function $g(\cdot)$ (with its inverse $g^{-1}(\cdot)$) exists that maps the conditional expectation $\mathsf{E}[\boldsymbol{y} \mid \boldsymbol{\gamma}]$ linearly on the $n \times (p+1)$ predictor matrix $\boldsymbol{X}$ with its corresponding $(p+1) \times 1$ fixed effect vector $\boldsymbol{\beta}$ and on the $n \times tG$ block-predictor matrix $\boldsymbol{Z} = (\boldsymbol{Z}_1 \quad \cdots \quad \boldsymbol{Z}_G)$ with its corresponding $tG \times 1$ random-effect vector $\boldsymbol{\gamma}$. (Here, t are the number of assumed random-effect predictors, and G is the number of units that the random effects should vary across; see below.) A third and final assumption is that the random effects follow some form of distribution and not necessarily a normal one.

The dependent variables in LSA studies are usually the plausible values, which are typically treated as continuous and, in multilevel analyses, as normally distributed (Atar and Atar 2012; Boulifa and Kaaouachi 2015; Caponera and Losito 2016; Cosgrove and Cunningham 2011; Leino and Malin 2006; Lüdtke *et al.* 2008; Martin *et al.* 2013; Mohammadpour 2013; Mohammadpour *et al.* 2015; OECD 2016a,b; Smith *et al.* 2016; Sun *et al.* 2012; Wendt *et al.* 2017). In terms of the GLMM, this treatment implies the need to use the identity link function with the normal distribution as the conditional density. Hence,

$$\mathsf{E}[\boldsymbol{y} \mid \boldsymbol{\gamma}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma},$$

[8] For the sake of simplicity here and in the following text, we suppress the additional conditional elements $\boldsymbol{\beta}$ and $\sigma^2_{y_i}$ in the expression of $\mathsf{E}[y_i \mid \boldsymbol{\gamma}]$ and $f(y_i \mid \boldsymbol{\gamma})$.

[9] Usually, and hereafter, the assumption is that $f(y_i \mid \boldsymbol{\gamma})$ belongs to the exponential family or is similar to the exponential family. In the expression for this distribution, $h(y_i)$ is the base, η_i is the natural parameter of the respective exponential distribution, $T(y_i)$ is the sufficient statistic of $f(y_i \mid \boldsymbol{\gamma})$,$A(\cdot)$ is a log-partition function of the natural parameter η_i, and ϕ is a dispersion parameter.

with the conditional probability density

$$f(y_i \mid \boldsymbol{\gamma}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{\frac{-(y_i-\mu_i)^2}{2\sigma^2}}.$$

Instead of using the plausible values, Gilleece, Cosgrove, and Sofroniou (2010) used the proficiency values z_j^m for these variables as the dependent variable when conducting multilevel analyses. The proficiency values can be considered as ordinal variables (see above), and the GLMM can accommodate these by assuming, for example, the multinomial distribution as the conditional density of the data and by using the cumulative logit function as the link function. Thus, assuming y_i can fall in $m = 1, \ldots, M$ categories, then

$$\mathsf{E}[y_i \leq j \mid \boldsymbol{\gamma}] = \mathsf{P}(y_i \leq j) = \nu_{ij} = \frac{e^{\delta_j - \lambda_i}}{1 + e^{\delta_j - \lambda_i}} \quad j = 1, \ldots, M-1,$$

where ν_{ij} are the cumulative probabilities $\nu_{ij} = p_{i1} + \cdots + p_{ij}$ (with p_{ij} as the probability that observation i will fall into category j), δ_j are the intercept for category j, and $\lambda_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}$. The conditional probability density function for the multinomial is

$$f(y_{i1}, \cdots, y_{iM} \mid \boldsymbol{\gamma}) = \frac{m!}{y_{i1}! \cdots y_{iM}!} p_{i1} \cdots p_{iM}.$$

In addition to using the proficiency values, researchers sometimes use a binary variable as the dependent variable when conducting multilevel analyses with LSA datasets (Karakolidis, Pitsia, and Emvalotis 2016; Rabe-Hesketh and Skrondal 2006; Zhu 2014). The binary variable is usually a function of the plausible values. Thus, for example, students with a scale score not above the low international benchmark will be assigned a zero and the remaining students will be assigned a one on the binary variable. The logistic link function is often used as a means of accommodating binary responses in the GLMM. Thus, if we assume $y_i = 0$ when observation i belongs to category 1 and $y_i = 1$ when observation i belongs to category 2, then the link function of the GLMM can be written as

$$\mathsf{E}[y_i = 1 \mid \boldsymbol{\gamma}] = \mathsf{P}(y_i = 1) = p_i = \frac{e^{\lambda_i}}{1 + e^{\lambda_i}},$$

with $\lambda_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}$. The corresponding conditional density for this model is

$$f(y_i \mid \boldsymbol{\gamma}) = \frac{e^{-\left(\frac{y_i-\mu_i}{s}\right)}}{s\left(1 + e^{-\left(\frac{y_i-\mu_i}{s}\right)}\right)^2},$$

where $s = 1$ is the most common default. Of course, the GLMM can accommodate many other link functions and families of the conditional density (McCulloch *et al.* 2008). We provided the above examples because they seem to us the most important ones in the context of LSA studies.

The GLMM can also accommodate the multilevel structure of LSA datasets through proper specification of $\boldsymbol{Z}$ and $\boldsymbol{\gamma}$, and we can demonstrate this possibility by using the example of a two-stage sampling design with a continuous normally distributed response variable. Extending this example to models with more than two stages or with response variables

that are not normally distributed is straightforward. Let us, then, assume a two-stage cluster sampling design where schools were randomly sampled at stage 1 and students were randomly sampled at stage 2. Let $j = 1, \ldots, G$ and $i = 1, \ldots, n_j$ denote the indices of the units at level 2 (stage 1) and level 1 (stage 2). Assume also that mathematical achievement is the response variable y_{ij}, and that students' social status x_{1ij} and school region (urban or rural) x_{2ij} explain students' achievement in mathematics. If we further assume that the school-based average achievement of students and the relationship between social status and mathematical achievement vary randomly across schools, then we can assume that the result will be a random intercept and a random slope model. In terms of the GLMM, this model can be expressed as

$$\begin{aligned}\mathsf{E}[\boldsymbol{y} \mid \boldsymbol{\gamma}] &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}, \\ &= \begin{pmatrix} 1 & x_{111} & x_{211} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n_11} & x_{2n_11} \\ 1 & x_{112} & x_{212} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n_G1} & x_{2n_G1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & x_{111} & & & \\ \vdots & \vdots & & & \\ 1 & x_{1n_11} & & & \\ & & \ddots & & \\ & & & 1 & x_{11G} \\ & & & \vdots & \vdots \\ & & & 1 & x_{1n_GG} \end{pmatrix} \begin{pmatrix} \gamma_{01} \\ \gamma_{11} \\ \vdots \\ \gamma_{0G} \\ \gamma_{1G} \end{pmatrix}.\end{aligned}$$

where $\gamma_{01}, \ldots, \gamma_{0G}$ are the random effects due to the intercept, and $\gamma_{11}, \ldots, \gamma_{1G}$ are the random effects due to the slope of social status (with the zeros in $\boldsymbol{Z}$ replaced by empty spaces). Because we are unlikely to be interested in the predicted values of $\boldsymbol{\gamma}_j = (\gamma_{0j} \quad \cdots \quad \gamma_{tj})^\top$ but instead interested in the variances and covariance of these effects, our goal with respect to our example would typically be that of estimating

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_j} = \begin{pmatrix} \sigma^2_{\gamma_0} & \sigma_{\gamma_0,\gamma_1} \\ \sigma_{\gamma_1,\gamma_0} & \sigma^2_{\gamma_1} \end{pmatrix},$$

rather than predicting $\boldsymbol{\gamma}_j$ directly. Finally, let us assume that $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_j} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ holds for all j (for a critical reflection on this assumption, see Daniels and Zhao 2003; Heagerty and Zeger 2000).

In the following text we denote the probability of a single observation, conditional on the fixed and random effects, by $p(y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_j)$.[10] We can then express the likelihood of the data having normally distributed random effects $\boldsymbol{\gamma}_j \sim N_G(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ as

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_G) = \prod_{j=1}^{G} L_j(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}_j) = \prod_{j=1}^{G} p(\boldsymbol{y}_j \mid \boldsymbol{\beta}),$$

with the $n_j \times 1$ response vectors $\boldsymbol{y}_j = (y_{1j} \quad \cdots \quad y_{n_jj})^\top$, and

$$p(\boldsymbol{y}_j \mid \boldsymbol{\beta}) = \int \prod_{i=1}^{n_j} p(y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_j) \varphi(\boldsymbol{\gamma}_j \mid \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) d\boldsymbol{\gamma}_j = \int f(\boldsymbol{y}_j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_j) \varphi(\boldsymbol{\gamma}_j \mid \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) d\boldsymbol{\gamma}_j, \tag{5}$$

where $\varphi(\boldsymbol{\gamma}_j \mid \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ is the multivariate normal distribution of dimension $v = t + 1$ (when a random intercept is assumed) or $v = t$ (without a random intercept). In some GLMMs,

[10] In the Bernoulli case considered above, for example, $p(y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_j) = p_{ij}^{y_{ij}}(1 - p_{ij})^{1-p_{ij}}$.

the integral in Equation 5 has a closed-form solution, as occurs, for example, when the dependent variable is normally distributed and the linear mixed model is consequently used. However, there are also models for which no analytic solution to this integral exists. The literature contains many different proposed methods for estimating the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_\gamma$ in these cases (Booth and Hobert 1999; Breslow and Clayton 1993; Breslow and Lin 1995; Gamerman 1997; Lin and Breslow 1996; Natarajan and Kass 2000; Pinheiro and Bates 1995; Rabe-Hesketh and Skrondal 2006; Raudenbush, Yang, and Yosef 2000; Shun 1997; Shun and McCullagh 1995; Wolfinger and O'Connell 1993; Zeger and Karim 1991). According to Tuerlinckx *et al.* (2006) and the SAS Institute Inc. (2020), these methods can be represented by two general types of solution. The first approximates the integral numerically (Booth and Hobert 1999; Gamerman 1997; Natarajan and Kass 2000; Rabe-Hesketh and Skrondal 2006; Zeger and Karim 1991) and the second approximates the integrand (Breslow and Clayton 1993; Breslow and Lin 1995; Lin and Breslow 1996; Raudenbush *et al.* 2000; Shun 1997; Shun and McCullagh 1995; Wolfinger and O'Connell 1993), which means that the integral of this approximation has a closed form. Our macro **SURVEYHLM** uses maximum quasi-likelihood as the estimation method (e.g., for approximating the integrand) for models without random effects and for multilevel models with a normally distributed dependent variable when level-specific weights are not specified, whereas the adaptive Gaussian quadrature is used (e.g., for approximating the integral) for all other models.

2.5. The SURVEYHLM macro

The **SURVEYHLM** macro fits multilevel models (Dempster, Rubin, and Tsutakawa 1981; Hofmann 1997; Lindley and Smith 1972; Rabe-Hesketh and Skrondal 2006; Raudenbush and Bryk 2002; Smith 1973; Woltman *et al.* 2012) with LSA datasets. It produces descriptive statistics for the variables in the model, can handle up to three levels, and makes it possible to specify, separately for each level, a random intercept, random slopes (Raudenbush and Bryk 2002), or both. It can also assume different correlation structures for the random components, and it fits both general linear mixed models and linear mixed models (McCulloch *et al.* 2008). It furthermore allows plausible values to be used as response variables, the response does not necessarily need to be normally distributed, and users can specify both continuous and categorical independent variables for each level. Centering the predictors around the class mean is possible, as is centering around the grand mean (Raudenbush and Bryk 2002).

Weights can be specified on levels 1, 2, and 3, and the combined weight can also be specified. If a combined weight is specified, the descriptive statistics are based on this weight; if not, the descriptive statistics are unweighted. Various researchers recommend using level-specific weights with LSA datasets (Asparouhov 2006; Pfeffermann *et al.* 1998; Rabe-Hesketh and Skrondal 2006; Rutkowski *et al.* 2010). Consequently, if a multilevel analysis defines level-specific weights, the **SURVEYHLM** macro uses these weights as the default in multilevel analyses. If there is no such definition, the macro uses the combined weight or performs an unweighted analysis. The **SURVEYHLM** macro also allows the level-specific weights to be scaled (Asparouhov 2006; Carle 2009; Pfeffermann *et al.* 1998; Rabe-Hesketh and Skrondal 2006).

The standard errors for both descriptive statistics and multilevel coefficients can be based on a sandwich estimator or calculated with the jackknife replication technique or through the provision of user-supplied replication weights (Grilli and Pratesi 2004; Kolenikov 2010; Korn

and Graubard 2003; Kovacevic, Rong, and You 2006; Lohr 2010; Rust and Rao 1996; Wolter 2007). The macro's sandwich estimators of the standard errors for the multilevel coefficients are based on the classical empirical-based estimator (White 1982), and the macro supports either a jackknife replication procedure or user-supplied replication weights (see Section 2). If the user uses plausible values as dependent variables and calculates standard errors with either jackknife or user-supplied replication weights, then he or she can decide whether to use all plausible values or only the first plausible value for standard error calculations (see Section 2).

We need, at this point, to make a comment about estimating of standard errors with the jackknife replication technique or through user-supplied replication weights. For descriptive statistics such as a function of totals, the properties of replication variance estimation techniques are well studied (Kolenikov 2010; Lohr 2010; Rust and Rao 1996; Wolter 2007) and typically result in consistent estimators of the standard errors. For multilevel models with multilevel weights, various researchers have proposed different resampling approaches to estimating the standard errors (Korn and Graubard 2003; Kovacevic *et al.* 2006). In addition, some software packages contain resampling approaches that include level-specific weights (see Table 1). However, little is yet known about how proposed and implemented estimation techniques behave within the context of multilevel models with multilevel weights. Korn and Graubard (2003) investigate a jackknife estimator with joint inclusion probabilities in a simulation study that involved simple random sampling and reported reasonable estimators of the standard errors. Unfortunately, the authors did not provide further details of the study's design and its results. Grilli and Pratesi (2004) conducted a simulation study that featured a complex sample design to investigate, among other considerations, Korn and Graubard's (2003) proposed jackknife estimator. The authors claimed that the proposed estimator appeared to be unreliable in this setting. However, it is difficult to assess the veracity of their claim, as they did not detail their results. In a study that used informative sampling, Kovacevic *et al.* (2006) assessed the behavior of two different bootstrapping estimators and a sandwich estimator of the standard errors for multilevel models with multilevel weights. In general, given the relative bias, it seems that bootstrap variance estimators overestimate and sandwich estimators underestimate the variance. However, it also seems that across all methods the bias decreases somewhat when the sample size increases (an outcome perhaps indicative of consistency). However, the authors simulated only four sample-size combinations and provided only graphical presentations of their results. They also gave no indication of whether these results were statistically significant. As such, a comprehensive study designed to investigate and compare the different suggested and implemented methods for estimating standard errors in multilevel models with multilevel weights is still needed, while further theoretical and numerical research directed toward the behavior of these estimators remains desirable.

The estimation method the macro uses for fixed effects models (e.g., general linear models or linear models) and for multilevel models with a normally distributed dependent variable when level-specific weights are not specified is the maximum pseudo-likelihood method (Breslow and Clayton 1993; Shall 1991; Tuerlinckx *et al.* 2006; Wolfinger and O'Connell 1993). The method used for the maximum likelihood estimates in all other models is the adaptive Gauss-Hermit quadrature method (Pinheiro and Bates 1995; Pinheiro and Chao 2006; Raudenbush *et al.* 2000; Tuerlinckx *et al.* 2006; Wolfinger and O'Connell 1993). The default optimizer is the trust-region method (TRUREG). However, users can also use one of several alternative opti-

mizers, namely, the Newton-Raphson method with ridging (NRRIDG), the Newton-Raphson method with line-search (NEWRAP), the quasi-Newton method (QUANEW), the double-dogleg method (DBLDOG), the conjugate gradient method (CONGRA), and the Nelder-Mead simplex method (NMSIMP; Fletcher 2001). Finally, the macro enables specification of the maximum number of iterations and function calls. Details about the **SURVEYHLM** macro code are set out in Appendix A.

3. Comparison of available software

Research on multilevel-analysis models, especially the GLMM, has not only flourished since the early 1980s but also been accompanied by the development of a variety of software programs. Because it is beyond the scope of this paper to compare all these programs, we restrict our comparison to (a) the packages for multilevel or GLMM analysis that are available in R (R Core Team 2025), and (b) the programs most frequently used for multilevel analysis of LSA datasets (see Section 2.4).[11] The second set of programs includes HLM (Raudenbush, Bryk, and Congdon 2013), Mplus(Muthén and Muthén 2017), Stata (StataCorp 2019), and SAS/STAT (SAS Institute Inc. 2022), and the first set (i.e., those programs are most often used in R) are **glmm** (Knudson 2017), **nlme** (Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2017), **lme4** (Bates, Mächler, Bolker, and Walker 2015), **MASS** (Venables and Ripley 2002), **glmmTMB** (Brooks *et al.* 2017), **glmmLasso** (Groll and Tutz 2014; Groll 2017), **GLMMadaptive** (Rizopoulos 2020), and **MCMCglmm** (Hadfield 2010).[12]

The `glmm` function in the package **glmm** fits the GLMM by using the ordinary Monte Carlo (MC) method to approximate the likelihood. Available distributions include the binomial, the Bernoulli, and the Poisson distribution. Multiple random components (i.e., multiple levels) are possible. However, the package assumes that $\boldsymbol{G}$ ($= \boldsymbol{\Sigma}_\gamma$) is diagonal wherever distinct variance components can be set as equal. The `lme` function in the package **nlme** provides more flexibility in terms of defining $\boldsymbol{G}$, and it also fits linear mixed-effects models. The maximum likelihood (ML) approach or the restricted maximum likelihood (REML) approach can be used for the estimation. Multiple random components are possible, as are flexible definitions of the structure of $\boldsymbol{G}$ (e.g., unstructured or diagonal). The packages also allow for modeling of the structures for the (level 1) residual matrix $\boldsymbol{R}$ (e.g., heteroscedasticity), but the Gaussian default distributional family with the identity link cannot be changed. This change is possible with the `glmer` function of the package **lme4**. Available distributions include, for example, the binomial, the Poisson, and the gamma distribution. Depending on the assumed distribution, the link function can be, for example, the logit link or the probit link, while the Laplace approximation (LA) or adaptive Gauss-Hermit quadrature (AGH) can be used for maximum likelihood estimation. Multiple random components are possible, but the AGH is not available for multiple random components. Also, although the **lme4** package provides some functionality for modeling the structure of $\boldsymbol{G}$, it does not provide the same flexibility for defining this structure that the **nlme** package does. In addition, defining a structure for $\boldsymbol{R}$ (other than

[11] For example, software such as the **Mixed-Up** Suite (Hedeker and Gibbons 1996a,b), the **ASReml** program (Gilmour, Gogel, Cullis, Welham, and Thompson 2015), the R packages **rstan** (Carpenter *et al.* 2017), **PLmixed** (Rockwood and Jeon 2019), **blme** (Chung, Rabe-Hesketh, Dorie, Gelman, and Liu 2013), and **multilevel** (Bliese 2016) are not part of our comparison. In general, these packages contain no features additional to those covered by the software considered here. Also, to the best of our knowledge, educational research typically does not use these packages to analyze LSA datasets.

[12] For an introduction to multilevel analysis with R, see Finch, Bolin, and Kelley (2014).

homoscedasticity) is not currently possible. The `glmmPQL` function in the **MASS** package uses penalized quasi-likelihood (PQL) to fit GLMMs. It offers different distributional families and link functions as well as ability to define the structure of $\boldsymbol{R}$. Multiple random components are possible, but the software assumes that the random effects are i.i.d. The **glmmTMB** package fits GLMM using LA estimation via the **TMB** package (Kristensen, Nielsen, Berg, Skaug, and Bell 2016). It is especially useful when zero-inflated count data should be analyzed, because it includes the Conway-Maxwell-Poisson distribution for the dependent variable. Other distributions, for example, the binomial, the Poisson, and the Gaussian, can also be specified, of course. Flexible definitions of the structure of $\boldsymbol{G}$ are possible and, in general, more than three levels can be estimated. The **glmmLasso** package also fits GLMM using LA estimation. However, a penalty term is included in the corresponding log-likelihood function allowing for automatic variable selection by L1-penalized estimation. Hence, this package is particularly helpful when a set of relevant independent variables should be selected from a set of many independent variables. Available distributions include the Gaussian, the binomial, and the Poisson. Flexible definitions of the structure of $\boldsymbol{G}$ are possible, and, in general, more than three levels can be estimated. With the **GLMMadaptive** package, GLMM for non-normally distributed dependent variables can be fitted. Available distributions include the binomial or the Poisson. AGH is used as the estimation method, and $\boldsymbol{G}$ can be either unstructured or diagonal. However, the **GLMMadaptive** package cannot be used to estimate models that have more than two levels. The `MCMCglmm` function of the **MCMCglmm** package enables Bayesian estimation of the GLMM. It also allows multiple responses to follow different types of distribution. The `MCMCglmm` furthermore allows for multiple random effects, and for definition of residual $\boldsymbol{R}$ and random-effect $\boldsymbol{G}$ variance structures.

A major drawback of all these packages with respect to analyzing LSA datasets is that none of them considers the special features of these datasets (i.e., plausible values, weighting, repeated replication techniques). However, the `withReplicates` function in the **survey** package (Lumley 2004, 2016) does, of course, enable estimation of the replication-based sampling variance s^2, while the function `withPV` enable the use of plausible values. It should be mentioned here, that the **intsvy** package (Caro and Biecek 2017) does consider the special features of LSA datasets. To our knowledge, it is the most comprehensive R package for analyzing LSA datasets. It provides tools for importing, merging, and analyzing data from international assessment studies. Among the available analyses functions are mean statistics, standard deviations, regression estimates, correlation coefficients, and frequency tables. However, mixed model estimation in general and GLMM analyses in particular are not supported by this package.

We are reasonably certain that the only multilevel-analysis R function that simultaneously accounts for the special features of LSA datasets is the `BIFIE.twolevelreg` function from the **BIFIEsurvey** package (BIFIE 2017). This function uses full maximum likelihood (FML) estimation to fit two-level linear mixed-effects models. Level-specific weights are feasible with this package, as is definition of multiple random components. Constraining elements of $\boldsymbol{G}$ to fixed values is also possible. The standard errors either can be based on a sandwich estimator or can be computed using the repeated replication technique, while the `BIFIE.data` function or the `BIFIE.data.jack` function can be used to define the survey sample design. However, estimates of the sample variance are based on the M repeated application of Equation 4 (one cycle for each plausible value) as described in step 10 of Section 2.3 of this article (see also Bruneforth, Oberwimmer, and Robitzsch 2016). Hence, if we wanted to analyze TIMSS and

PIRLS data in the manner it was usually done until 2011 (i.e., with estimates of the sample variance based only on the first plausible value), then we would have to perform the desired analysis (e.g., sample mean, multilevel analysis) separately for each plausible value and then merge the results manually using the post-processing functionality of R. Furthermore, we would not be able to change the identity link function or the default Gaussian distribution.

Version 7 of the HLM program (Raudenbush *et al.* 2013) fits hierarchical models with up to four levels. Available distributions include the Gaussian, the binomial, the Poisson, and the multinomial. Depending on the assumed distribution and the number of levels, researchers can use LA, AGH, REML, or FML as the estimation method. Multiple random components are possible, and level-specific weights can be defined. Heterogeneous residual variances are also possible while, in general, $\boldsymbol{G}$ is assumed to be unstructured. The program supports the use of plausible values, but does not permit repeated replication estimation. Version 8 of Mplus (Muthén and Muthén 2017), among other such programs, enables researchers to implement hierarchical models of up to three levels. The dependent variable for two-level models can be continuous, censored, binary, ordinal, nominal, or counts, and for three-level models continuous or categorical. Available estimators include FML (up to three levels), limited information weighted least squares (LLS; Asparouhov and Muthén 2007), the Muthén limited information estimator (MLS; Muthén 1994), and the Bayes estimator (up to three levels). Multiple random components are possible, and level-specific weights can be used. Mplus also offers considerable flexibility in defining the structure of both $\boldsymbol{R}$ and $\boldsymbol{G}$. The program furthermore supports plausible values, but does not permit user-supplied replication weights (for multilevel analysis). Stata (StataCorp 2019) provides many functions for fitting multilevel models. Of these, the most commonly used are the `meglm` and `gllamm` (Rabe-Hesketh, Skrondal, and Pickles 2004, 2005). The `meglm` function fits multilevel mixed-effects GLMs. Available distributions include the Gaussian, binomial, gamma, and Poisson. LA, two versions of AGH, and nonadaptive Gauss-Hermit quadrature (GH) can be used as the estimator. Multiple random components are possible with Stata. Level-specific weights can be used and the structure of $\boldsymbol{G}$ defined. The `gllamm` function fits generalized linear latent and mixed models. Available distributions include the Gaussian, binomial, gamma, and Poisson. AGH or nonadaptive GH can be used as the estimation method. Multiple random components are possible, level-specific weights are supported, and $\boldsymbol{G}$ is assumed to be unstructured. Although these two functions do not provide options for handling plausible values or producing replication variance, the `pv` function (designed specifically for PISA, TIMSS, and PIRLS student achievement data; Macdonald 2008) and `repest` (designed to be used with PISA and other OECD study datasets; Avvisati and Keslair 2014) enable estimation with weighted replicate samples and plausible values. However, the `pv`'s default estimation method for TIMSS and PIRLS data corresponds to the procedure typically used in TIMSS and PIRLS until 2011. Analyses focused on TIMSS 2015 or PIRLS 2016 datasets may therefore not provide the estimates for the standard errors that are usually reported.

SAS/STAT provides different procedures for estimating multilevel models. Among those used most often are `PROC MIXED`, `PROC NLMIXED`, and `PROC GLIMMIX`. The first of these uses either the REML, ML, or MIVQUE0 (Hartley, Rao, and LaMotte 1978) method to fit linear mixed models. Multiple random components are possible, and the structure of $\boldsymbol{R}$ and $\boldsymbol{G}$ can be defined flexibly. However, the procedure does not permit use of level-specific weights. `PROC NLMIXED`, however, does support level-specific weights in models with no more than two levels. This procedure uses AGH, GH, or the first-order method (TS) developed by colleagues Beal

and Sheiner (1982, 1988) and Shiner and Beal (1985) to fit nonlinear mixed-effects models. The procedure allows for multiple random components and enables flexible definition of the structure of $\boldsymbol{G}$ and $\boldsymbol{R}$. `PROC NLMIXED` likewise provides users with a very flexible means of defining the conditional distribution of the data (given the random effects). However, the standard forms of this distribution (e.g., Gaussian, binary, binomial, gamma, Poisson) are also available. The `PROC GLIMMIX` procedure fits generalized linear mixed models. LA, AGH, GH, or different pseudo-likelihood (PL) techniques can be used, with choice of technique dependent on the model. Multiple random components are possible, level-specific weights are supported, and the structure of $\boldsymbol{G}$ and $\boldsymbol{R}$ can be defined very flexibly. The built-in conditional distributions of the data include the Gaussian, binary, binomial, gamma, Poisson, and multinomial. Users can also define their own mean and variance function. In short, `PROC GLIMMIX` is a highly general procedure. However, like the other SAS/STAT procedures considered here, `PROC GLIMMIX` does not consider the special features of LSA datasets, which again explains why we developed the SAS **SURVEYHLM** macro. The macro is based on the `PROC GLIMMIX` procedure, but we expanded it so that it can also handle plausible values and repeated replication techniques.

SAS contains not only procedures for estimating multilevel models but also procedures specifically designed for survey sample designs: `PROC SURVEYSELECT`, `PROC SURVEYFREQ`, `PROC SURVEYMEANS`, `PROC SURVEYREG`, `PROC SURVEYLOGISTIC`, `PROC SURVEYPHREG` and `PROC SURVEYIMPUTE`. With `PROC SURVEYSELECT` a variety of methods are available for selecting probability-based (simple or complex multistage) random samples. In order to calculate one-way to n-way frequency and cross tabulation tables from sample survey data `PROC SURVEYFREQ` can be used. `PROC SURVEYFREQ` (like the other survey procedures) can handle data from complex multistage survey designs with stratification, clustering, and unequal weighting. It provides a choice of variance estimation methods (for example, bootstrap or jackknife). `PROC SURVEYMEANS` computes statistics such as means, totals, proportions, quantiles, geometric means, and ratios from a survey sample. `PROC SURVEYREG` performs linear regression analysis for sample survey data, `PROC SURVEYLOGISTIC` fits linear logistic regression models for discrete response survey data, and `PROC SURVEYPHREG` performs regression analysis based on the Cox proportional hazard model for sample survey data. Finally, `PROC SURVEYIMPUTE` uses a fractional hot-desk imputation method or some other traditional hot-desk imputation technique to impute missing values in a dataset. The procedure also creates replicate weights that account for the imputation and that can be used for replication-based variance estimation for complex surveys. However, none of these procedures can handle plausible values. Hence, if plausible values are used with these procedures then a procedure like `PROC MIANALYZE` must be used to combine the results obtained for each plausible value separately. In addition, these procedures do not support GLMM analyses.

Table 1 presents a summative comparison of the software discussed in this section of our article (the SAS survey procedures discussed above were not included in the table because they do not support GLMM analyses). Compared to the other GLMM procedures already implemented in SAS, the **SURVEYHLM** macro can handle plausible values. Also, because the user can supply replication weights, the **SURVEYHLM** macro provides a variety of variance estimation methods based on the repeated replication technique (in addition to a sandwich estimator of the standard errors). Level-specific weights are supported by the macro, and different scaling procedures are available to scale these weights. Moreover, because the macro supports level-specific replication weights, it can be used to estimate resampling approaches

Software	Function	Features										
		Estimation	Family		Levels			$\boldsymbol{G}$	$\boldsymbol{R}$	PVs	Level-specific weights	RRT
			Gaussian	Other	2	3	>3					
R	`glmm`	MC	✓	✓	✓	✓	✓	d		✓[a]		✓[b]
	`lme`	ML, REML	✓		✓	✓	✓	f	f	✓[a]		✓[a]
	`glmer`	LA, AGH	✓	✓	✓	✓	✓	f		✓[a]		✓[b]
	`glmmPQL`	PQL	✓	✓	✓	✓	✓	d	f	✓[a]		✓[b]
	`glmmTMB`	LA	✓	✓	✓	✓	✓	f		✓[a]		✓[b]
	`glmmLasso`	LA	✓	✓	✓	✓	✓	f		✓[a]		✓[b]
	`GLMMadaptive`	AGH		✓	✓			f		✓[a]		✓[b]
	`MCMCglmm`	MCMC	✓	✓	✓	✓	✓	f	f	✓[a]		✓[b]
	`BIFIE.twolevelreg`	FML	✓		✓			f		✓	✓	✓[c]
HLM		LA, AGH REML, FML	✓	✓	✓	✓	✓	u		✓	✓	
Mplus		FML, LLS MLS, Bayes	✓	✓	✓	✓		f	f	✓	✓	
Stata	`meglm`	LA, AGH, GH	✓	✓	✓	✓	✓	f		✓[d]	✓	✓[d]
	`gllamm`	AGH, GH	✓	✓	✓	✓	✓	u		✓[d]	✓	✓[d]
SAS	`PROC MIXED`	REML, ML MIVQUE0	✓		✓	✓	✓	f	f			
	`PROC NLMIXED`	AGH, GH, TS	✓	✓	✓	✓	✓	f	f		✓[e]	
	`PROC GLIMMIX`	LA, AGH GH, PL	✓	✓	✓	✓	✓	f	f		✓	
	SURVEYHLM	AGH, PL	✓	✓	✓	✓		f		✓	✓	✓

Estimation abbreviations: PVs, plausible values; RRT, repeated replication technique; MC, Monte Carlo; ML, maximum likelihood; REML, restricted maximum likelihood; LA, Laplace approximation; AGH, adaptive Gauss-Hermit quadrature; PQL, penalized quasi-likelihood; MCMC, Markov chain Monte Carlo; FML, full maximum likelihood; LLS, limited information weighted least square; MLS, Muthén limited information estimator; GH, nonadaptive Gauss-Hermit quadrature; TS, first-order method; PL, pseudo-likelihood techniques. $\boldsymbol{G}$ abbreviations: d, diagonal; f, flexible; u, unstructured. Repeated replication technique notes: [a]Uses the `withPV` function of the **survey** package. [b]Uses the `withReplicates` function of the **survey** package. [c]The method implemented for TIMSS and PIRLS corresponds to the method used in TIMSS and PIRLS since 2015. [d]Uses the `pv` or `repest` function (the implemented method for TIMSS and PIRLS corresponds to the method used in TIMSS and PIRLS until 2011). [e]Supports level-specific weights for models with no more than two levels.

Table 1: Comparison of available software for multilevel analyses.

that include level-specific weights. Centering the predictors around the class mean is possible, as is centering around the grand mean. In addition to the results of the GLMM analysis our macro can be used to produce descriptive statistics for LSA datasets. Hence, the **SURVEYHLM** macro can be used, for example, to fit not only a three level linear mixed model with random intercepts and random slopes on either level two, level three, or both levels, with plausible values as the dependent variable (see Section 5.1) but also a three level multinomial mixed model using the benchmark values (TIMSS and PIRLS) or the proficiency levels (PISA) as the dependent variable (see Section 5.2). The standard errors for both of these analyses can be based on a sandwich estimator or on different repeated replication techniques. Because the macro allows for the specification of different covariance structures for the variance components, users can also test hypotheses about specific covariance structures. For example, users can test the hypothesis that the variances and covariances of multiple random slopes are equal or that the random slopes do not correlate at all (see Section 5.3). Finally, in order to conduct a diagnostic check of the GLMM analysis, the macro can print out different residual plots and ensure that during any further analyses the estimated fixed and random effects for each dependent variable (e.g., plausible value) and the combined model are saved by default.

4. Invocation of the SURVEYHLM macro

The following set of code invokes the **SURVEYHLM** macro. Necessary arguments are written in uppercase letters (`NEST3` is a necessary argument only if a three-level model is analyzed). Any default values are written after the equals sign.

```
%macro surveyhlm(DATN = , ROOTPV = , NPV = 1, noint = n, xvar1 = , xvar2 = ,
    xvar3 = , cvar1 = , cvar2 = , cvar3 = , ccent2 = , ccent3 = , gcent = ,
    norint2 = n, norint3 = n, rslope2 = , rslope3 = , NEST1 = , NEST2 = ,
    NEST3 = , l1wgt = , l2wgt = , l3wgt = , wgt = , sfw = 1, sfb2 = 2,
    sfb3 = 2, jkrep = , jkzone = , nrwgt = , jktyp = full, jkfac = 0.5,
    repwp = , shrtcut = n,  srvysam = y, odesc = n, graph = n, label = model,
    qpoints = , tec = trureg, gconv = 1E-8, maxfunc = , maxiter = , type2 = ,
    type3 = , ldata2 = , clist2 = , ldata3 = , clist3 = , dist = normal,
    start = y, startrw = y, libd = , libe = );
```

The macro parameters are as follows.

- `DATN`: The input SAS dataset to be used for the analysis.
- `ROOTPV`: The dependent variable of the model. If `&NPV = 1`, it signifies the name of the dependent variable. If `&NPV > 1`, then `ROOTPV` is the root name of the dependent variable. The names of the plausible values in the dataset must therefore be `ROOTPV&N`, with `N` denoting an index that ranges from 1 to `&NPV`.
- `NPV`: If plausible values are used, then the number of plausible values must be specified here, else one.
- `noint`: If no intercept is required in the fixed effects model, then `noint = y`, else `noint = n`.

- `xvar1`, `xvar2`, and `xvar3`: The variable names of the continuous predictors in the multilevel model for levels 1, 2, and 3.

- `cvar1`, `cvar2`, and `cvar3`: The variable names of the categorical predictors in the multilevel model for levels 1, 2, and 3.

- `ccent2`: The names of the level 1 (continuous) predictor variables that should be centered around the level 2 means (with the class mean centering on level 2). The variable names must also appear in `xvar1` (or in `cvar1`).

- `ccent3`: The names of the level 1 or level 2 (continuous) predictor variables that should be centered around the level 3 means (with the class mean centering on level 3). The variable names must also appear in `xvar1` or `xvar2` (or in `cvar1` and `cvar2`).

- `gcent`: The names of the (continuous) predictor variables that should be centered around the grand mean (i.e., the grand mean centering). The variable names must also appear in `xvar1`, `xvar2`, or `xvar3` (or in `cvar1`, `cvar2`, and `cvar3`).

- `norint2` and `norint3`: If a random intercept is to be included on level 2 (level 3), then `norint2 = n` (`norint3 = n`), else `norint2 = y` (`norint3 = y`).

- `rslope2`: The names of the level 1 predictors with random slopes on level 2. The variable names must also appear in `xvar1` or in `cvar1`.

- `rslope3`: The names of the level 1 or level 2 predictors with random slopes on level 3. The variable names must also appear in `xvar1`, `xvar2`, `cvar1`, or `cvar2`.

- `NEST1`: The name of the level 1 identifier variable. This is usually an ID assigned to students, for example, IDSTUD in PIRLS and TIMSS.

- `NEST2`: The name of the level 2 identifier variable. This is usually an ID assigned to classes (or schools), for example, IDCLASS (or IDSCHOOL) in PIRLS and TIMSS.

- `NEST3`: The name of the level 3 identifier variable. This is usually an ID assigned to schools (or countries), for example, IDSCHOOL (or IDCNTRY) in PIRLS and TIMSS.

- `l1wgt`, `l2wgt`, and `l3wgt`: The name of the variable with the level 1, level 2, and/or level 3 specific weight.

- `wgt`: The name of the variable with the combined weight, for example, HOUWGT in PIRLS and TIMSS.

- `sfw`: Scaling of the level-1 specific weight `l1wgt` (0: unscaled; 1: scaled to sample size; 2: scaled to effective sample size).

- `sfb2`: Scaling of the level-2 specific weight `l2wgt`, where 0: unscaled; 1: sum of `l1wgt * l2wgt` is the sample size (with `l1wgt` unscaled); 2: sum of `l1wgt * l2wgt` is the sample size (with `l1wgt` scaled to sample size); and 3: sum of `l1wgt * l2wgt` is the effective sample size (with `l1wgt` scaled to the effective sample size).

- `sfb3`: Scaling of the level-3 specific weight `l3wgt`, where 0: unscaled; 1: sum of `l1wgt * l3wgt` is the sample size (with `l1wgt` unscaled); 2: sum of `l1wgt * l3wgt` is the sample size (with `l1wgt` scaled to sample size); and 3: sum of `l1wgt * l3wgt` is the effective sample size (with `l1wgt` scaled to the effective sample size).

- `jkrep`: The name of the variable with the jackknife replication code, for example, JKREP in PIRLS and TIMSS.

- `jkzone`: The name of the variable with the jackknife zone assignment, for example, JKZONE in PIRLS and TIMSS.

- `nrwgt`: The number of replication samples that should be used, with `nrwgt = 150` having usually been used for PIRLS and TIMSS since 2015 and `nrwgt = 75` for PIRLS and TIMSS usually used before 2015.

- `jktyp`: The type for constructing the jackknife replication weights (see Section 2), with `jktyp = full` having usually been used for PIRLS and TIMSS since 2015 and `jktyp = half` usually used for PIRLS and TIMSS before 2015.

- `jkfac`: The variance factor for the replication variance, with `jkfac = 0.5` having usually been used for PIRLS and TIMSS since 2015 and `jkfac = 1.0` usually used for PIRLS and TIMSS before 2015.

- `repwp`: The root name of the user-supplied replication weights, which means `repwp&i` must be the names of the user-supplied replication variables in the dataset.

- `shrtcut`: If the replication variance should be the average across all plausible values, then `shrtcut = n`; else if the replication variance should be based only on the first plausible value, then `shrtcut = y`, with `shrtcut = n` having usually been used for PIRLS and TIMSS since 2015 and `shrtcut = y` usually used for PIRLS and TIMSS before 2015.

- `srvysam`: If standard errors of the multilevel fixed and random effects should be based on the replication technique, then `srvysam = y`; else if these standard errors should be based on the sandwich estimator, then `srvysam = n`.

- `odesc`: If only descriptive statistics should be estimated and no multilevel analysis should be performed, then `odesc = y`, else `odesc = n`.

- `graph`: Should residual plots be printed, then `graph = y`, else `graph = n`.

- `label`: The label for the analysis.

- `qpoints`: The number of quadrature points in each dimension of the integral for fitting the random-effect models. If not specified (the default), the number of quadrature points is selected adaptive. If there are v random effects for each subject and `qpoints = n`, then n^v evaluations (or $(npv)n^v$ if plausible values are used) of the conditional log likelihood for each observation are necessary to compute one value of the objective function. Increasing the number of quadrature nodes can therefore substantially increase the computational burden. This outcome is especially likely if the standard errors of the multilevel parameter estimates should be based on the replication technique. This

is because the number of evaluations increases to $(nrwgt + 1)n^v$ (or $(nrwgt + npv)n^v$ if plausible values are used. When `qpoints = 1`, the adaptive quadrature approximation results are similar to the results of the Laplace approximation.

- `tec`: Determines the optimization technique. Possible values are TRUREG, NRRIDG, NEWRAP, QUANEW, DBLDOG, CONGRA, and NMSIMP.

- `gconv`: Specifies a relative gradient convergence criterion (see the documentation for the `nloptions` statement for the `PROC GLIMMIX` procedure in SAS Institute Inc. 2017).

- `maxfunc`: Specifies the maximum number of function calls in the optimization process. The default value depends on the optimization technique (see the documentation for the `nloptions` statement for the `PROC GLIMMIX` procedure in SAS Institute Inc. 2017).

- `maxiter`: Specifies the maximum number of iterations in the optimization process. The default value depends on the optimization technique (see the documentation for the `nloptions` statement for the `PROC GLIMMIX` procedure in SAS Institute Inc. 2017).

- `type2` and `type3`: Specifies the structure of $\boldsymbol{R}$ for fixed effects models and the covariance structure $\boldsymbol{G}$ for random effects models on level 2 and level 3 (for details, see the `random` statement for the `PROC GLIMMIX` procedure in SAS Institute Inc. 2017).

- `ldata2` and `ldata3`: If `type2 = lin(q)` (`type3 = lin(q)`), then `ldata2` (`ldata3`) is a SAS dataset with the matrices of the assumed linear combinations.

- `clist2` and `clist3`: If `type2 = sp(exp)(c-list)` (or `type3 = sp(exp)(c-list)`), then `clist2` (`clist3`) is a list of variable names for the argument `(c-list)`.

- `dist`: Specifies the (conditional) probability distribution of `ROOTPV` (see the documentation for the `model` statement for the `PROC GLIMMIX` procedure in SAS Institute Inc. 2017).

- `start`: If plausible values are used as dependent variables, then the random components estimated by using the first plausible value can be used as starting values for estimating these components for the other plausible values (`start = y`); else if `start = n`, then the built-in procedure for estimating the starting values for the random components is used for all plausible values (for details, see the `PROC GLIMMIX` procedure in SAS Institute Inc. 2017).

- `startrw`: If the user requests replication-based standard errors for the multilevel parameter estimates, then estimates of the random components for the model with the full sample can be used as starting values for the estimates of these components when replicated samples are used. Thus, `startrw = y`; else `startrw = n` if the built-in procedure for estimating the starting values for the random components is used for all replications (for details, see the `PROC GLIMMIX` procedure in SAS Institute Inc. 2017).

- `libd`: A libref for the dataset `DATN`. The default is the working directory.

- `libe`: A libref where the results of the analysis should be saved. The default is the working directory.

5. Application of the SURVEYHLM macro

This section includes three examples of using the macro for analyses of PIRLS and TIMSS data. We chose these examples because they allowed us to account for a variety of analysis specifications, such as continuous and categorical dependent and independent variables, centering, two-level and three-level analyses, and handling TIMSS' and PIRLS' data from before and after 2011 (see Section 2 of this article). Example 3 also accommodates the specific testing of assumptions on the variance and covariance structures within the model.

5.1. Application 1: Two-level versus three-level analysis

The first example specifies a multilevel analysis on two versus three levels. The dependent variable (`asmmat0`) is the continuous student mathematics achievement scale from TIMSS 2015. Predictors include the motivation to learn mathematics at the individual student level, that is, level 1 (`SLM_sca0`), teachers' confidence in their ability to adapt their teaching at the class level, that is, level 2 (`AT_MEAN`), and the overall use of the language of the LSA test within schools at the class level versus the school level, that is, level 3 (`ACBG04d`). Conceptually, the difference modeled at level 2 is that the language used in the test is assumed to be a class characteristic, yet the language used in the whole school is likely to be the same throughout the school if all teachers in it are teaching the school curriculum in that one language. If this is the case, then the language characteristic can be modeled on level 3. The data that we used for our analysis (`T15_prep`) was the mathematics achievement data from the Dutch fourth-grade students who participated in TIMSS 2015 (Mullis *et al.* 2016a,b).

As shown by, for example, Deci and Ryan (1985), students' intrinsic motivation is a strong predictor of their achievement in a school subject, mainly because the extent to which they are interested in that subject tends to determine the extent to which they engage in learning it and thus achieving in it (Pintrich, Smith, Garcia, and McKeachie 1991). Mathematics is no exception in this regard, as is evident from the TIMSS 2015 data, where the achievement of Dutch students who said they liked learning mathematics very much was, on average, 38 scale score points higher than the average achievement of students who said they did not like learning maths (Mullis *et al.* 2016a). We therefore included, in our analysis, motivation as a level-1 predictor of student achievement in mathematics.

At the class level, adapted teaching (i.e., teaching directed toward meeting students' individual interests and learning needs) has become a focus of interest among educational stakeholders in recent years due to the increase in student heterogeneity in many classrooms worldwide. Adapted teaching has been associated with fostering student achievement because it makes the subject more relevant to the individual learner (see, for example, Schulz-Heidorf and Solheim 2016; Van De Pol, Volman, Oort, and Beishuizen 2014). Our analysis therefore modeled the influence that teachers' confidence in adapting their teaching had on the TIMSS 2015 mathematics scores of the fourth-grade Dutch students. Because the extent to which students are familiar (native) users of the language of a test is known to have a strong influence on their achievement on that test (Mullis *et al.* 2017), our analysis tested whether this variable showed as a composite effect when considered as a school (versus class) indicator.

For $i = 1, \ldots, n_j$ students in $j = 1, \ldots, G$ classes, the model that we estimate correspond to the following equation

$$\texttt{asmmat0}_{ij} = \beta_0 + \texttt{SLM\_sca0}_{ij} \times \beta_1 + \texttt{ACBG04d}_j \times \beta_2 + \texttt{AT\_MEAN}_j \times \beta_3 + \gamma_{0j} + \epsilon_{ij},$$

Convergent status

Convergent	ASMMAT01	ASMMAT02	ASMMAT03	ASMMAT04	ASMMAT05
Status	0	0	0	0	0
Reason	Convergence criterion (GCONV=1E-8) satisfied.	Convergence criterion (GCONV=1E-8) satisfied.	Convergence criterion (GCONV=1E-8) satisfied.	Convergence criterion (GCONV=1E-8) satisfied.	Convergence criterion (ABSGCONV=0.00001) satisfied.

Number of observations

Label	ASMMAT01	ASMMAT02	ASMMAT03	ASMMAT04	ASMMAT05
Number of Observations Read	4515	4515	4515	4515	4515
Number of Observations Used	2459	2459	2459	2459	2459

Fit statistics

Description	Value
-2 Log Likelihood	26499,87
AIC (smaller is better)	26513,87
AICC (smaller is better)	26513,92
BIC (smaller is better)	26533,56
CAIC (smaller is better)	26540,56
HQIC (smaller is better)	26521,87

Conditional fit statistics

Description	Value
-2 log L(ASMMAT05 | r. effects)	26195,17
Pearson Chi-Square	6073694,98
Pearson Chi-Square / DF	2469,99

Random components (covariances)

Effect	Row	Col1
Intercept	1	189,38

Random components (correlations)

Effect	Row	COL1
Intercept	1	1

Random components (standard errors)

CovParm	Subject	Estimate	StdErr
UN(1,1)	IDCLASS	189,378541	49,105
Residual		2686,585172	158,554

Fixed effects

NImpute	Parm	SLM_sca0	ACBG04d	Estimate	StdErr	DF	tvalue	Probt
5	Intercept	.	.	555,321152	9,9674	293,48	55,7139	<.0001
5	AT_MEAN	.	.	-1,838213	3,1832	605,35	-0,5775	0,2819
5	SLM_sca0	0	.	-29,846283	3,5438	911,83	-8,422	<.0001
5	SLM_sca0	1	.	-16,480691	3,1891	436,61	-5,1679	<.0001
5	SLM_sca0	2	.	0	0	.	.	.
5	ACBG04d	.	0	-31,458616	16,106	1772,23	-1,9532	0,0255
5	ACBG04d	.	1	0	0	.	.	.

Figure 1: Output for the two-level analysis obtained with the **SURVEYHLM** macro.

where γ_{0j} is the random intercept term. The syntax that we used for estimating this model was as follows.

```
%surveyhlm(DATN = T15_prep, ROOTPV = asmmat0, NPV = 5,  CVAR1 = SLM_sca0,
    CVAR2 = ACBG04d, XVAR2 = AT_MEAN, GCENT = at_mean, NEST1 = idstud,
    NEST2 = idclass, WGT = MATWGT, L1WGT = WGT_L1, L2WGT = WGT_L2o,
    JKREP = JKREP, JKZONE = JKZONE, NRWGT = 150, LABEL = model12Lsyv,
    TYPE2 = UN, QPOINTS = 1, LIBD = &dpath, LIBE = &dpath);
```

Thus, for the dataset `T15_prep`, we analyzed a two-level model with a student identifier for level 1 (`NEST1 = idstud`) and a class identifier for level 2 (`NEST2 = idclass`), with five plausible values (`NPV = 5`) serving as dependent variables. Because the dataset came from

TIMSS 2015, we were able to use `JKREP` and `JKZONE` together with `NRWGT = 150` and the default values `JKTYP = full`, and `JKFAC = 0.5` to build appropriate replication weights. We requested an unstructured form for the variance components (`TYPE2 = UN`).

Figure 1 depicts the output for the two-level analysis. First, the convergent status for the multilevel analysis for each of the five plausible values is reported. As can be seen, all five models fulfill the convergence criterion. Therefore, it can be assumed that all five models converged. Second, the number of observations read and the number of observations used for the analysis are reported. Overall, the dataset includes $n = 4,515$ records, with $n = 2,459$ of these cases used for the analysis (the drop out is due to missing values on at least one of the variables). While in the first two tables the results are depicted separately for each of the five plausible values, the point estimates of the following tables are based on the average across the five plausible values according to Rubin's formula (Section 2.1), and the standard errors are based on the repeated replication technique (Section 2.3). Consequently, for example, the different fit statistics depicted in the next table of the output are the average across the fit statistics for the separate plausible values.

The fit statistics are particularly useful for comparing competing models. However, because these fit statistics depend on the random components, the conditional fit statistics are also given. The estimated value of the random component ($\hat{\sigma}_I^2 = 189.38$), the corresponding standard error ($\hat{\sigma}_{SE.I} = 49.11$), and the residual variance ($\hat{\sigma}_e^2 = 2,686.59$) are printed next. In general, given these results, assuming a model with a random intercept is plausible. In the final table of the output estimates for the fixed effects are reported. For each fixed effect, the number of imputations, the parameter name, the level of the variable (for non-continuous independent variables), the estimated value, the standard error of the estimate, the degrees of freedom, the t-value, and the corresponding probability value are given. In this two-level model the effects for the first dummy (0 = *very much like learning maths*; 1 = *do not like learning maths*) and second dummy (0 = *very much like learning maths*; 1 = *like learning maths*) of `SLM_sca0` as well as for `ACBG04d` (0 = *speaks language of test*; 1 = *do not speak language of test*) are significant. Hence, given this two-level model, one would assume that the motivation to learn mathematics and that the test language is the same as the student's native language are positively associated with maths achievement.

For the three-level analysis, we used basically the same syntax as the syntax we used for the two-level model. However, instead of having just a two-level identifier we now, of course, needed a three-level identifier (i.e., `NEST1 = idstud`, `NEST2 = idclass`, and `NEST3 = idschool`), corresponding level-3 specific weights (i.e., `L1WGT = WGT_L1`, `L2WGT = WGT_L2`, `L3WGT = WGT_L3`), and the independent variable `ACBG04d` set on level 3.

For $i = 1, \ldots, n_j$ students in $j = 1, \ldots, G_k$ classes and $k = 1, \ldots, K$ schools, the model that we estimated corresponded to the following equation

$$\begin{aligned} \texttt{asmmat0}_{ijk} = \beta_0 + \texttt{SLM\_sca0}_{ijk} \times \beta_1 + \texttt{ACBG04d}_k \times \beta_2 + \texttt{AT\_MEAN}_j \times \beta_3 \\ + \gamma_{0j} + \gamma_{0k} + \epsilon_{ijk}, \end{aligned}$$

where γ_{0j} is the random intercept term for class membership and γ_{0k} is the random intercept term for school membership. The syntax that we used for this model was as follows.

```
%surveyhlm(DATN = T15_prep, ROOTPV = asmmat0, NPV = 5, CVAR1 = SLM_sca0,
    CVAR3 = ACBG04d, XVAR2 = AT_MEAN, GCENT = at_mean, NEST1 = idstud,
```

Convergent status

Convergent	ASMMAT01	ASMMAT02	ASMMAT03	ASMMAT04	ASMMAT05
Status	0	0	0	0	0
Reason	Convergence criterion (GCONV=1E-8) satisfied.	Convergence criterion (GCONV=1E-8) satisfied.	Convergence criterion (GCONV=1E-8) satisfied.	Convergence criterion (ABSGCONV=0.0 0001) satisfied.	Convergence criterion (GCONV=1E-8) satisfied.

Number of observations

Label	ASMMAT01	ASMMAT02	ASMMAT03	ASMMAT04	ASMMAT05
Number of Observations Read	4515	4515	4515	4515	4515
Number of Observations Used	2459	2459	2459	2459	2459

Fit statistics

Description	Value
-2 Log Likelihood	26419,06
AIC (smaller is better)	26435,06
AICC (smaller is better)	26435,12
BIC (smaller is better)	26454,41
CAIC (smaller is better)	26462,41
HQIC (smaller is better)	26442,83

Conditional fit statistics

Description	Value
-2 log L(ASMMAT05 | r. effects)	26200,61
Pearson Chi-Square	6090918,6
Pearson Chi-Square / DF	2476,99

Random components (covariances)

Stmt	Effect	Subject	Row	Col1	Col2	Col3	Col4
1	Intercept	IDCLASS(IDSCHOOL) 301 3	1	95,2428	0	0	0
1	Intercept		2	0	95,2428	0	0
1	Intercept		3	0	0	95,2428	0
1	Intercept		4	0	0	0	95,2428
2	Intercept	IDSCHOOL 3	1	91,2332	.	.	.

Random components (correlations)

Stmt	Effect	Subject	Row	COL1	COL2	COL3	COL4
1	Intercept	IDCLASS(IDSCHOOL) 301 3	1	1	0	0	0
1	Intercept		2	0	1	0	0
1	Intercept		3	0	0	1	0
1	Intercept		4	0	0	0	1
2	Intercept	IDSCHOOL 3	1	1	.	.	.

Random components (standard errors)

CovParm	Subject	Estimate	StdErr
UN(1,1)	IDCLASS(IDSCHOOL)	95,242819	78,119
UN(1,1)	IDSCHOOL	91,233181	68,799
Residual		2688,806395	158,749

Fixed effects

NImpute	Parm	SLM_sca0	ACBG04d	Estimate	StdErr	DF	tvalue	Probt
5	Intercept	.	.	552,16745	10,0079	426,18	55,1734	<.0001
5	AT_MEAN	.	.	-0,929822	3,2007	664,11	-0,2905	0,3858
5	SLM_sca0	0	.	-29,893138	3,5382	1086,7	-8,4486	<.0001
5	SLM_sca0	1	.	-16,591164	3,1977	375,93	-5,1884	<.0001
5	SLM_sca0	2	.	0	0	.	.	.
5	ACBG04d	.	0	-28,937033	16,0697	406,16	-1,8007	0,0362
5	ACBG04d	.	1	0	0	.	.	.

Figure 2: Output for the three-level analysis obtained with the **SURVEYHLM** macro.

```
NEST2 = idclass, nest3 = idschool, WGT = MATWGT, L1WGT = WGT_L1,
L2WGT = WGT_L2, L3wgt = WGT_L3, JKREP = JKREP, JKZONE = JKZONE,
NRWGT = 150, LABEL = model13Lsyv, TYPE2 = UN, TYPE3 = UN, QPOINTS = 1,
startrw = n, LIBD = &dpath, LIBE = &dpath);
```

Figure 2 depicts the output for the three-level analysis. As can be seen, compared to the two-level analysis, the estimated random intercept variance on level two has decreased to $\hat{\sigma}^2_{I.L2} = 95.24$, while the random intercept variance on level three is now $\hat{\sigma}^2_{I.L3} = 91.23$. In addition, all fit statistics for the three-level model are smaller than the corresponding fit statistics for the two-level model, suggesting that the three-level model is more suitable for these data than the two-level model. With respect to the fixed effects, in the three-level model the significant factors are the same as the significant factors in the two-level model.

5.2. Application 2: Multilevel analysis with categorical dependent variables

We chose a categorical dependent variable for our second analysis. This time, we used PIRLS 2011 data from the German students who participated in the study (Foy and Drucker 2013; Mullis, Martin, Foy, and Drucker 2012). PIRLS categorized students' reading literacy achievement scores into benchmarks, with scores of 400 points and under marking the low international benchmark, scores between 401 and 475 points marking the intermediate international benchmark, scores between 478 and 550 points denoting the high international benchmark, and scores between 551 and 625 points or more denoting the advanced international benchmark (Foy and Drucker 2013, see also Section 2.1 of this article). When conducting our second analysis, we categorized the dependent variable (`bmhr0`) as 0 = *students at the high and advanced benchmarks* and 1 = *students at the intermediate and low benchmarks*.

As the predictor of student performance at the individual level, PIRLS used the number of books in the student's home (`bookr` with 1 = *0-200 books* and 0 = *more than 200 books*) as an indicator of a family's cultural capital (Bourdieu 1983; Gustafsson, Hansen, and Rosén 2013). The number of books at home is seen as an indicator of the extent to which a family values education and provides a supportive learning environment, which, in turn, influences how well students perform at school (Bradley and Corwyn 2002; Noble, McCandless, and Farah 2007).

A student's performance on a reading test can potentially also be explained by teacher characteristics such as formal training (Myrberg 2007). While most of the questions on this matter in the PIRLS 2011 teacher survey focused on pedagogical or didactical aspects, such as teaching reading, educational psychology, special education, assessment methods in reading, and the like, the question on the extent to which teachers had studied reading theory tapped into a more theoretical aspect of reading literacy that nonetheless might influence how many students in a teacher's class reach the advanced or high international reading literacy benchmarks. We investigated this consideration using the variable `readtr` as our predictor of student achievement (per benchmark) in our two-level model. Here, 1 = *reading theory was not studied, or teachers had an overview or introduction to the topic* and 0 = *reading theory was an area of emphasis*.

For $i = 1, \ldots, n_j$ students in $j = 1, \ldots, G$ schools, the model that we estimated corresponded to the following equation

$$\mathsf{E}[\texttt{bmhr0}_{ij} = 0 \mid \boldsymbol{\gamma}_j] = \mathsf{P}(\texttt{bmhr0}_{ij} = 0) = \frac{e^{\lambda_{ij}}}{1 + e^{\lambda_{ij}}},$$

with $\lambda_{ij} = \beta_0 + bookr_{ij} \times \beta_1 + readtr_j \times \beta_2 + \gamma_{0j}$, where γ_{0j} is the random intercept term for school membership. The syntax that we used for this analysis was as follows. Note, however, in comparison to our first analysis, that we needed to specify the distribution parameter `DIST = binary`.

```
%surveyhlm(DATN = p11_model2, ROOTPV = bmhr0, NPV = 5, CVAR1 = bookr,
    cvar2 = readtr, NEST1 = idstud, NEST2 = idschool, WGT = TCHWGT,
    L1WGT = studwgt, L2WGT = classwgt, JKREP = JKREP, JKZONE = JKZONE,
    NRWGT = 75, JKTYP = half, JKFAC = 1.0, SHRTCUT = y, LABEL = modell2syv,
    TYPE2 = UN, dist = BINARY, startrw = n, QPOINTS = 7, LIBD = &dpath,
    LIBE = &dpath);
```

Fixed effects

NImpute	Parm	bookr	readtr	Estimate	StdErr	DF	tvalue	Probt
5	Intercept	.	.	0,527949	0,10995	305,624	4,8017	<.0001
5	bookr	0	.	0	0,10614	.	0	.
5	bookr	1	.	-0,820161	0,04893	164,062	-16,7617	<.0001
5	readtr	.	0	0,109259	0,16253	242,508	0,6722	0,251
5	readtr	.	1	0	0	.	.	.

Figure 3: Output for the multilevel analysis with categorical dependent variable in **SURVEYHLM** (**SURVEYHLM** models the probability of $\mathsf{P}(y = 0)$ for binary dependent variables when `DIST = binary`).

As can be seen in Figure 3, this second model indicated that students from homes with fewer than 200 books were less likely than the students with more than 200 books in their home to attain the advanced or high international benchmarks. The emphasis on reading theory during teachers' formal training, however, showed no association with students' benchmark placements.

5.3. Application 3: Testing for specific covariance structures

For our third model, we decided to conduct an analysis that tested structures beyond those in the typical multilevel analysis. Because the `TYPE`-statement allows for a variety of covariance structures (for details, see the `random` statement of the `PROC GLIMMIX` procedure in SAS Institute Inc. 2017), we selected statement `TYPE = TOEP(2)`. It specifies a two-banded Toeplitz covariance matrix that assumes equal variances in the slopes and equal covariance of the random terms. It also assumes that one of the three predictors will not correlating with the remaining two. In essence, this model is useful for analyses that use an effect-coded ordinal variable as the predictor variable. The assumption is that the variation of the regression slopes (the differences between the current code and the reference group) is the same for all steps involving the ordinal variable. In addition, because covariates exist only between the adjacent steps (equal), it is they, the adjacent steps, that show relationships. We again used German PIRLS data from 2011 (Foy and Drucker 2013; Mullis *et al.* 2012) to specify this structure in the **SURVEYHLM** macro. Our dependent variable was overall reading achievement with an average achievement score of 541 score points. Our predictors of students' overall reading achievement were students' availability of books at home (`books`, with 0 = *0–10 books*, 1 = *11–25 books*, 2 = *26–100 books*, 3 = *101–200 books*, and 4 = *more than 200 books*). The association between number of books at home and academic achievement is generally positive (Mullis *et al.* 2017).

While the regression slopes of such structures are tested relatively often in multilevel analyses, complex structures involving slope variances and the covariances of their random terms have seldom been tested to date. To analyze whether different numbers of books at home showed equal variance in the random effects across classes, and equal covariance between adjacent steps in the ordinal-coded variable, we used the **SURVEYHLM** macro with the specification `TYPE2 = TOEP(2)`.

For $i = 1, \ldots, n_j$ students in $j = 1, \ldots, G$ schools, the model that we estimated corresponded

Random components (correlations)								
Effect	**books**	**Row**	**COL1**	**COL2**	**COL3**	**COL4**	**COL5**	**COL6**
Intercept	_	1	1	0,55215	0	0	0	0
books	none or few (0-10)	2	0,55215	1	0,55215	0	0	0
books	one bookcase (26-100)	3	0	0,55215	1	0,55215	0	0
books	one shelf (11-25)	4	0	0	0,55215	1	0,55215	0
books	three or more bookcases (200+)	5	0	0	0	0,55215	1	0,55215
books	two bookcases (101-200)	6	0	0	0	0	0,55215	1

Figure 4: Random components (correlations) with a two-banded Toeplitz covariance structure in the **SURVEYHLM** macro.

to the following equation

$$\begin{aligned}\texttt{asrrea0}_{ij} = \beta_0 + \texttt{books}_{1.ij} \times \beta_1 + \texttt{books}_{2.ij} \times \beta_2 + \texttt{books}_{3.ij} \times \beta_3 + \texttt{books}_{4.ij} \times \beta_4 + \gamma_{0j} \\ + \texttt{books}_{1.ij} \times \gamma_{1j} + \texttt{books}_{2.ij} \times \gamma_{2j} + \texttt{books}_{3.ij} \times \gamma_{3j} + \texttt{books}_{4.ij} \times \gamma_{4j} \\ + \texttt{books}_{5.ij} \times \gamma_{5j} + \epsilon_{ij},\end{aligned}$$

where γ_{0j} is the random intercept term for the school membership and $\gamma_{1j}, \ldots, \gamma_{5j}$ are the random slope terms for the dummies of the `books` variable. For the variance components in $\boldsymbol{\Sigma}_{\gamma_j}$ we assume

$$\boldsymbol{\Sigma}_{\gamma_j} = \begin{pmatrix} \sigma^2 & \sigma_1 & 0 & 0 & 0 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_1 & \sigma^2 & \sigma_1 & 0 & 0 \\ 0 & 0 & \sigma_1 & \sigma^2 & \sigma_1 & 0 \\ 0 & 0 & 0 & \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & 0 & 0 & 0 & \sigma_1 & \sigma^2 \end{pmatrix},$$

when `TYPE2 = TOEP(2)` and

$$\boldsymbol{\Sigma}_{\gamma_j} = \begin{pmatrix} \sigma^2_{\gamma_0} & \sigma_{\gamma_0,\gamma_1} & \sigma_{\gamma_0,\gamma_2} & \sigma_{\gamma_0,\gamma_3} & \sigma_{\gamma_0,\gamma_4} & \sigma_{\gamma_0,\gamma_5} \\ \sigma_{\gamma_1,\gamma_0} & \sigma^2_{\gamma_1} & \sigma_{\gamma_1,\gamma_2} & \sigma_{\gamma_1,\gamma_3} & \sigma_{\gamma_1,\gamma_4} & \sigma_{\gamma_1,\gamma_5} \\ \sigma_{\gamma_2,\gamma_0} & \sigma_{\gamma_2,\gamma_1} & \sigma^2_{\gamma_2} & \sigma_{\gamma_2,\gamma_3} & \sigma_{\gamma_2,\gamma_4} & \sigma_{\gamma_2,\gamma_5} \\ \sigma_{\gamma_3,\gamma_0} & \sigma_{\gamma_3,\gamma_1} & \sigma_{\gamma_3,\gamma_2} & \sigma^2_{\gamma_3} & \sigma_{\gamma_3,\gamma_4} & \sigma_{\gamma_3,\gamma_5} \\ \sigma_{\gamma_4,\gamma_0} & \sigma_{\gamma_4,\gamma_1} & \sigma_{\gamma_4,\gamma_2} & \sigma_{\gamma_4,\gamma_3} & \sigma^2_{\gamma_4} & \sigma_{\gamma_4,\gamma_5} \\ \sigma_{\gamma_5,\gamma_0} & \sigma_{\gamma_5,\gamma_1} & \sigma_{\gamma_5,\gamma_2} & \sigma_{\gamma_5,\gamma_3} & \sigma_{\gamma_5,\gamma_4} & \sigma^2_{\gamma_5} \end{pmatrix},$$

when `TYPE2 = UN`. The syntax for this analysis was as follows.

```
%surveyhlm(DATN = P11_model3, ROOTPV = asrrea0, NPV = 5, CVAR1 = books,
    RSLOPE2 = books, NEST1 = idstud, NEST2 = idschool, WGT = TCHWGT,
    L1WGT = studwgt, L2WGT = classwgt, JKREP = JKREP, JKZONE = JKZONE,
    NRWGT = 75, JKTYP = half, JKFAC = 1.0, SHRTCUT = y, LABEL = model3toe,
    MAXFUNC = 1500, TYPE2 = TOEP(2), start = n, startrw = n, qpoints = 2,
    LIBD = &dpath, LIBE = &dpath);
```

As can be seen from the results in Figure 4, the correlations between the random effects of different numbers of books at home were estimated to be $r = 0.55$. Thus, the across-class correlation between the random slopes for different numbers of books was of medium

Random components (correlations)

Effect	books	Row	COL1	COL2	COL3	COL4	COL5	COL6
Intercept	_	1	1	-0,00015	-0,00053	-0,00034	-0,00006	-0,00028
books	none or few (0-10)	2	-0,00015	1	0,01984	0,06549	-0,1958	-0,16659
books	one bookcase (26-100)	3	-0,00053	0,01984	1	0,00051	-0,08828	-0,23491
books	one shelf (11-25)	4	-0,00034	0,06549	0,00051	1	0,14504	-0,13297
books	three or more bookcases (200+)	5	-0,00006	-0,1958	-0,08828	0,14504	1	-0,17683
books	two bookcases (101-200)	6	-0,00028	-0,16659	-0,23491	-0,13297	-0,17683	1

Figure 5: Random components (correlations) with an unstructured covariance structure in the **SURVEYHLM** macro.

strength when a two-banded Toeplitz covariance structure was assumed. In comparison, Figure 5, which presents the correlation matrix for an unstructured covariance matrix, shows correlations below (above) the second band, but most of them are very small. The fit statistics for these models (AIC = 83,427.2 and BIC = 83,528.6 for the unstructured case, and AIC = 83,066.2 and BIC = 83,096.2 for the two-banded Toeplitz case) also support the hypothesis of a two-banded Toeplitz covariance structure. While the outcome of this analysis warrants debate, it shows that the **SURVEYHLM** macro, along with the other presented specifications, allows for complex analyses beyond the scope of typical multilevel analyses of large-scale educational assessment data.

6. Conclusion

Based on a comparison of the software packages typically used to conduct multilevel analyses of large-scale assessment (LSA) datasets, we developed the SAS macro **SURVEYHLM** for researchers conducting multilevel analyses of large-scale educational assessment data. The **SURVEYHLM** macro fits multilevel models with LSA datasets and uses the GLMM for estimation purposes. The dependent variable can therefore be, for example, continuous plausible values or an ordinal transformation of these values (usually called benchmark values or proficiency levels). General linear models and linear models can also be fitted with this macro.

Maximum pseudo-likelihood is used as the estimation method for fixed effects models (Breslow and Clayton 1993; Shall 1991; Tuerlinckx *et al.* 2006; Wolfinger and O'Connell 1993), while maximum likelihood estimates with Gauss-Hermit quadrature are used for all other models (Pinheiro and Bates 1995; Pinheiro and Chao 2006; Raudenbush *et al.* 2000; Tuerlinckx *et al.* 2006; Wolfinger and O'Connell 1993). However, use of the Gauss-Hermit quadrature can slow down the estimation process, especially for models with more than three random effects. The software for v random effects and n quadrature points needs at least n_v (or $(npv)n^v$) evaluations of the conditional log likelihood for each observation. Because the current version of the **SURVEYHLM** macro does not support crossed random effects or repeated measures, it cannot be used to fit special kinds of multilevel models, such as item-response models.

We demonstrated, through three applications (analyses), the usefulness of the **SURVEYHLM** macro. These highlighted the possibilities that the macro presents for fitting three-level models with LSA datasets. These possibilities include estimating repeated replication technique-based standard errors, having response variables that are not normally distributed, and modeling different correlation structures for the random effect. Until now, no software package has been able to capture all these features simultaneously. Our applications also showed

that these possibilities lead to parameter estimators that are more in line with the technical requirements of LSA datasets. Because the other parameter estimators also lead to other interpretations of the results, future multilevel analyses of LSA datasets need to take these details into account.

Acknowledgments

The authors acknowledge the anonymous reviewers for the attention and expertise they generously shared to support the production of this article. We further thank Paula Wagemaker for pre-submission English editing support.

References

Adams RJ, Wu M (eds.) (2002). *PISA 2000 Technical Report*. OECD Publishing, Paris.

Akaike H (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, **AC-19**(6), 716–723. doi:10.1109/tac.1974.1100705.

Anderson CJ, Kim JS, Keller B (2014). "Multilevel Modeling of Categorical Response Variables." In P Lietz, J Cresswell, K Rust, RJ Adams (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, pp. 481–519. CRC Press, Boca Raton. doi:10.1201/b16061-27.

Asparouhov T (2006). "General Multi-Level Modeling with Sampling Weights." *Communications in Statistics – Theory and Methods*, **35**(3), 439–460. doi:10.1080/03610920500476598.

Asparouhov T, Muthén B (2007). "Computationally Efficient Estimation of Multilevel High-Dimensional Latent Variable Models." In *Proceedings of the 2007 Joint Statistical Meeting: ASA Section on Biometrics. Salt Lake City, UT*. URL https://www.statmodel.com/download/JSM2007000746.pdf.

Atar HY, Atar B (2012). "Investigating the Multilevel Effects of Several Variables on Turkish Students' Science Achievements on TIMSS." *Journal of Baltic Science Education*, **11**(2), 115–126. doi:10.33225/jbse/12.11.115.

Avvisati F, Keslair F (2014). "REPEST: Stata Module to Run Estimations with Weighted Replicate Samples and Plausible Values." Statistical Software Components, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s457918.html.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Beal SL, Sheiner LB (1982). "Estimating Population Kinetics." *CRC Critical Reviews in Biomedical Engineering*, **8**, 195–222. URL https://pubmed.ncbi.nlm.nih.gov/6754254/.

Beal SL, Sheiner LB (1988). "Heteroscedastic Nonlinear Regression." *Technometrics*, **30**, 327–338. doi:10.1080/00401706.1988.10488406.

Beaton AE, Johnson EG (1992). "Overview of the Scaling Methodology Used in the National Assessment." *Journal of Educational Measurement*, **26**(2), 163–175. doi:10.1111/j.1745-3984.1992.tb00372.x.

Berezner A, Adams RJ (2017). "Why Large-Scale Assessments Use Scaling and Item Response Theory." In P Lietz, JC Cresswell, KF Rust, RJ Adams (eds.), *Implementation of Large-Scale Education Assessments*, pp. 323–356. John Wiley & Sons, Chichester. doi:10.1002/9781118762462.ch13.

Bickel R (2007). *Multilevel Analysis for Applied Research: It's Just Regression!* Methodology in the Social Sciences. The Guilford Press, New York. doi:10.1080/10705510801922670.

BIFIE (2017). ***BIFIEsurvey**: Tools for Survey Statistics in Educational Assessment*. doi:10.32614/CRAN.package.bifiesurvey. R package version 2.3-18.

Birnbaum A (1968). "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In FM Lord, MR Novick (eds.), *Statistical Theories of Mental Test Scores*, pp. 397–479. Addison-Wesley Publishing, Reading.

Bliese P (2016). ***multilevel**: Multilevel Functions*. doi:10.32614/CRAN.package.multilevel. R package version 2.6.

Booth JG, Hobert JP (1999). "Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm." *Journal of the Royal Statistical Society B*, **61**(1), 265–285. doi:10.1111/1467-9868.00176.

Boulifa K, Kaaouachi A (2015). "The Relationship between the Home Resource for Learning and Science Achievement in TIMSS 2011: A Multilevel Analysis." *Applied Mathematical Sciences*, **9**(13), 637–652. doi:10.12988/ams.2015.48668.

Bourdieu P (1983). "Ökonomisches Kapital, Kulturelles Kapital, Soziales Kapital." In R Kreckel (ed.), *Soziale Ungleichheiten: Sonderband 2 der Zeitschrift Soziale Welt*, pp. 183–198. Schwarz, Göttingen.

Bradley RH, Corwyn RF (2002). "Socioeconomic Status and Child Development." *Annual Review of Psychology*, **53**(1), 371–399. doi:10.1146/annurev.psych.53.100901.135233.

Breslow NE, Clayton DG (1993). "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association*, **88**(421), 9–25. doi:10.1080/01621459.1993.10594284.

Breslow NE, Lin X (1995). "Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion." *Biometrika*, **82**(1), 81–91. doi:10.1093/biomet/82.1.81.

Brooks ME, Kristensen K, Van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Maechler M, Bolker BM (2017). "**glmmTMB** Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling." *The R Journal*, **9**(2), 378–400. doi:10.32614/rj-2017-066.

Bruneforth M, Oberwimmer K, Robitzsch A (2016). "Reporting Und Analysen [Reporting and Analysis]." In S Breit, C Schreiner (eds.), *Large-Scale Assessment Mit R: Methodische*

Grundlagen Der Österreichischen BildungsstandardÜberprÜfung, chapter 10, pp. 333–362. Facultas Verlags- und Buchhandels AG, Vienna.

Caponera E, Losito B (2016). "Context Factors and Student Achievement in the IEA Studies: Evidence from TIMSS." *Large-Scale Assessment in Education*, **4**(12), 1–22. doi:10.1186/s40536-016-0030-6.

Carle AC (2009). "Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations." *BMC Medical Research Methodology*, **9**(49), 1–13. doi:10.1186/1471-2288-9-49.

Caro DH, Biecek P (2017). "**intsvy**: An R Package for Analyzing International Large-Scale Assessment Data." *Journal of Statistical Software*, **81**(7), 1–44. doi:10.18637/jss.v081.i07.

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, **76**(1), 1–32. doi:10.18637/jss.v076.i01.

Cassell DL (2007). "Don't Be Loopy: Re-Sampling and Simulation the SAS Way." In *Proceedings of the SAS Global Forum 2007 Conference*, pp. 1–18. SAS Institute Inc., Cary. URL https://support.sas.com/resources/papers/proceedings/proceedings/forum2007/183-2007.pdf.

Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J (2013). "A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models." *Psychometrika*, **78**(4), 685–709. doi:10.1007/s11336-013-9328-2.

Cosgrove J, Cunningham R (2011). "A Multilevel Model of Science Achievement of Irish Students Participating in PISA 2006." *The Irish Journal of Education*, **39**, 57–73.

Daniels MJ, Zhao YD (2003). "Modelling the Random Effects Covariance Matrix in Longitudinal Data." *Statistics in Medicine*, **22**, 1631–1647. doi:10.1002/sim.1470.

De Leeuw J, Meijer E (eds.) (2008). *Handbook of Multilevel Analysis*. Springer-Verlag, New York.

Deci EL, Ryan RM (1985). *Intrinsic Motivation and Self-Determination in Human Behaviour*. Plenum, New York.

Demir İ, Kiliç S, Ünal H (2010). "Effects of Students' and Schools' Characteristics on Mathematics Achievement: Findings from PISA 2006." *Procedia – Social and Behavioral Sciences*, **2**(2), 3099–3103. doi:10.1016/j.sbspro.2010.03.472.

Dempster AP, Rubin DB, Tsutakawa RK (1981). "Estimation in Covariance Components Models." *Journal of the American Statistical Association*, **76**(374), 341–353. doi:10.1080/01621459.1981.10477653.

Ene M, Leighton EA, Blue GL, Bell BA (2015). *Multilevel Models for Categorical Data Using SAS PROC GLIMMIX: The Basics*. URL https://support.sas.com/resources/papers/proceedings15/3430-2015.pdf.

Finch WH, Bolin JE, Kelley K (2014). *Multilevel Modeling Using R*. Chapman & Hall/CRC, Boca Raton.

Fletcher R (2001). *Practical Methods of Optimization*. John Wiley & Sons, New York.

Foy P (2017). *TIMSS 2015 User Guide for the International Database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Foy P (2018). *PIRLS 2016 User Guide for the International Database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Foy P, Arora A, Stanco GM (2013). *TIMSS 2011 User Guide for the International Database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Foy P, Drucker KT (2013). *PIRLS 2011 User Guide for the International Database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Fraillon J, Schulz W, Ainley J (2013). *International Computer and Information Literacy Study Assessment Framework*. International Association for the Evaluation of Educational Achievement (IEA), Amsterdam.

Fraillon J, Schulz W, Friedman T, Ainley J, Gebhardt E (2015). *ICILS 2013 Technical Report*. International Association for the Evaluation of Educational Achievement (IEA), Amsterdam.

Gamerman D (1997). "Sampling from the Posterior Distribution in Generalized Linear Mixed Models." *Statistics and Computing*, **7**(1), 57–68. doi:10.1023/a:1018509429360.

Ghagar MNA, Othman R, Mohammadpour E (2011). "Multilevel Analysis of Achievement in Mathematics of Malaysian and Singaporean Students." *Journal of Educational Psychology and Counseling*, **2**, 285–304.

Gilleece L, Cosgrove J, Sofroniou N (2010). "Equity in Mathematics and Science Outcomes: Characteristics Associated with High and Low Achievement on PISA 2006 in Ireland." *International Journal of Science and Mathematics Education*, **8**, 475–496. doi:10.1007/s10763-010-9199-2.

Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R (2015). ***ASReml** Release 4.1. Functional Specification*.

Grilli L, Pennoni F, Rampichini C, Romeo I (2015). "Exploiting TIMSS and PIRLS Combined Data: Multivariate Multilevel Modelling of Student Achievement." Paper presented at the 47th Scientific Meeting of the Italian Statistical Society, Cagliari, Italy, URL https://arxiv.org/pdf/1409.2642v2.pdf.

Grilli L, Pratesi M (2004). "Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs." *Statistics Canada*, **30**(1), 93–103.

Groll A (2017). ***glmmLasso**: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*. doi:10.32614/CRAN.package.glmmlasso. R package version 1.5.1.

Groll A, Tutz G (2014). "Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation." *Statistics and Computing*, **24**(2), 137–154. doi:10.32614/CRAN.package.glmmlasso.

Gustafsson JE, Hansen KY, Rosén M (2013). "Effects of Home Background on Student Achievement in Reading, Mathematics, and Science at the Fourth Grade." In MO Martin, IVS Mullis (eds.), *TIMSS and PIRLS 2011: Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Grade—Implications for Early Learning*, pp. 181–287. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill. URL https://timssandpirls.bc.edu/timsspirls2011/international-database.html.

Hadfield JD (2010). "MCMC Methods for Multi-Response Generalized Linear Mixed Models: The **MCMCglmm** R Package." *Journal of Statistical Software*, **33**(2), 1–22. doi:10.18637/jss.v033.i02.

Hartley HO, Rao JNK, LaMotte L (1978). "A Simple Synthesis-Based Method of Variance Component Estimation." *Biometrics*, **34**, 233–244. doi:10.2307/2530013.

Heagerty PJ, Zeger SL (2000). "Marginalized Multilevel Models and Likelihood Inference." *Statistical Science*, **15**, 1–26. doi:10.1214/ss/1009212671.

Hedeker D, Gibbons RD (1996a). "**MIXOR**: A Computer Program for Mixed-Effects Ordinal Regression Analysis." *Computer Methods and Programs in Biomedicine*, **49**, 157–176. doi:10.1016/0169-2607(96)01720-8.

Hedeker D, Gibbons RD (1996b). "**MIXREG**: A Computer Program for Mixed-Effects Regression Analysis with Autocorrelated Errors." *Computer Methods and Programs in Biomedicine*, **49**, 229–252. doi:10.1016/0169-2607(96)01723-3.

Hofmann DA (1997). "An Overview of the Logic and Rationale of Hierarchical Linear Models." *Journal of Management*, **23**, 723–744. doi:10.1016/s0149-2063(97)90026-x.

Hox J (2010). *Multilevel Analysis: Techniques and Applications*. Quantitative Methodology Series, 2nd edition. Taylor & Francis, Mahwah.

Husén T, Postlethwaite TN (1996). "A Brief History of the International Association for the Evaluation of Educational Achievement (IEA)." *Assessment in Education*, **3**(2), 129–141. doi:10.1080/0969594960030202.

Ismail ME, Samsudin MA, Zain ANM (2014). "A Multilevel Study on Trends in Malaysian Secondary School Students' Science Attitude: Evidence from TIMSS 2011." *International Journal of Asian Social Science*, **4**(5), 572–584. doi:10.1002/sce.21028.

Johnson EG, Rust KF (1992). "Population Inferences and Variance Estimation for NAEP Data." *Journal of Educational Statistics*, **17**(2), 175–190. doi:10.3102/10769986017002175.

Judkins DR (1990). "Fay's Method for Variance Estimation." *Journal of Official Statistics*, **6**, 223–229.

Jung M, Carstens R (2015). *ICILS 2013 User Guide for the International Database*. International Association for the Evaluation of Educational Achievement (IEA), Amsterdam.

Karakolidis A, Pitsia V, Emvalotis A (2016). "Mathematics Low Achievement in Greece: A Multilevel Analysis of the Programme for International Student Assessment (PISA) 2012 Data." *Themes in Science & Technology Education*, **9**(1), 3–24.

Karim MR, Zeger SL (1992). "Generalized Linear Models with Random Effects: Salamander Mating Revisited." *Biometrics*, **48**(2), 631–644. doi:10.2307/2532317.

Knudson C (2017). ***glmm**: Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation*. doi:10.32614/CRAN.package.glmm. R package version 1.2.2.

Köhler H, Weber S, Brese F, Schulz W, Carstens R (2018). *ICCS 2016 User Guide for the International Database*. International Association for the Evaluation of Educational Achievement (IEA), Amsterdam.

Kolenikov S (2010). "Resampling Variance Estimation for Complex Survey Data." *Stata Journal*, **10**(2), 165–199. doi:10.1177/1536867x1001000201.

Korn EL, Graubard BI (2003). "Estimating Variance Components by Using Survey Data." *Journal of the Royal Statistical Society B*, **65**, 175–190. doi:10.1111/1467-9868.00379.

Kovacevic MS, Rong H, You Y (2006). "Bootstrapping for Variance Estimation in Multi-Level Models Fitted to Survey Data." In *ASA Proceedings of the Survey Research Methods Section*, pp. 3260–3269.

Kreft I, De Leeuw J (1998). *Introducing Multilevel Modeling*. Sage Publications, London.

Kristensen K, Nielsen A, Berg CW, Skaug H, Bell BM (2016). "**TMB**: Automatic Differentiation and Laplace Approximation." *Journal of Statistical Software*, **70**(5), 1–21. doi:10.18637/jss.v070.i05.

Kurada RR (2016). "Fitting Multilevel Hierarchical Mixed Models Using `PROC NLMIXED`." In *Proceedings of the SAS Global Forum 2016*. Las Vegas. URL https://support.sas.com/resources/papers/proceedings16/SAS4720-2016.pdf.

Leino K, Malin A (2006). "Could Confidence in ICT Boost Boys' Reading Performance?" In J Mejding, A Roe (eds.), *Northern Lights on PISA 2003: A Reflection from the Nordic Countries*, chapter 11, pp. 175–188. Nordic Council of Ministers, Copenhagen.

Lin X, Breslow NE (1996). "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion." *Journal of the American Statistical Association*, **91**(435), 1007–1016. doi:10.1080/01621459.1996.10476971.

Lindley DV, Smith AFM (1972). "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society B*, **34**(1), 1–41. doi:10.1111/j.2517-6161.1972.tb00885.x.

Liou PY, Hung YC (2015). "Statistical Techniques Utilized in Analyzing PISA and TIMSS Data in Science Education from 1996 to 2013: A Methodological Review." *International Journal of Science and Mathematics Education*, **13**, 1449–1468. doi:10.1007/s10763-014-9558-5.

Lohr S (2010). *Sampling: Design and Analysis*. Brooks/Cole, Boston.

Lüdtke O, Marsh HW, Robitzsch A, Trautwein U, Asparouhov T, Muthén B (2008). "The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies." *Psychological Methods*, **13**(3), 203–229. doi:10.1037/a0012869.

Lumley T (2004). "Analysis of Complex Survey Samples." *Journal of Statistical Software*, **9**(1), 1–19. doi:10.18637/jss.v009.i08. R package version 2.2.

Lumley T (2016). "**survey**: Analysis of Complex Survey Samples." R package version 3.32.

Macdonald K (2008). "PV: Stata Module to Perform Estimation with Plausible Values." Statistical Software Components, Boston College Department of Economics, Boston. URL https://ideas.repec.org/c/boc/bocode/s456951.html.

Martin MO, Foy P, Mullis IVS, O'Dwyer LM (2013). "Effective Schools in Reading, Mathematics, and Science at the Fourth Grade." In MO Martin, IVS Mullis (eds.), *TIMSS and PIRLS 2011: Relationships among Reading, Mathematics, and Science Achievement at the Fourth Grade – Implications for Early Learning*, chapter 3, pp. 109–178. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Martin MO, Mullis IVS (eds.) (2012). *Methods and Procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Martin MO, Mullis IVS, Hooper M (eds.) (2016). *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Martin MO, Mullis IVS, Hooper M (eds.) (2017). *Methods and Procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Masters G (2017). "The Science of Large-Scale Assessment." In P Lietz, JC Cresswell, KF Rust, RJ Adams (eds.), *Implementation of Large-Scale Educational Assessments*, pp. xvii–xix. John Wiley & Sons, Chichester.

Masters GN (1982). "A Rasch Model for Partial Credit Scoring." *Psychometrika*, **47**(2), 149–174. doi:10.1007/bf02296272.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman & Hall, London.

McCulloch CE, Searle SR, Neuhaus JM (2008). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.

Meroni EC, Vera-Toscano E, Costa P (2015). "Can Low Skill Teachers Make Good Students? Empirical Evidence from PIAAC and PISA." *Journal of Policy Modeling*, **37**, 308–323. doi:10.1016/j.jpolmod.2015.02.006.

Mislevy RJ (1991). "Randomization-Based Inference about Latent Variables from Complex Samples." *Psychometrika*, **56**(2), 177–196. doi:10.1007/bf02294457.

Mislevy RJ, Beaton A, Kaplan BA, Sheehan K (1992a). "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses." *Journal of Educational Measurement*, **29**(2), 133–161. doi:10.1111/j.1745-3984.1992.tb00371.x.

Mislevy RJ, Johnson EG, Muraki E (1992b). "Scaling Procedures in NAEP." *Journal of Educational Statistics*, **17**(2), 131–154. doi:10.3102/10769986017002131.

Mohammadpour E (2013). "A Three-Level Multilevel Analysis of Singaporean Eighth-Graders' Science Achievement." *Learning and Individual Differences*, **26**, 212–220. doi:10.1016/j.lindif.2012.12.005.

Mohammadpour E, Kalantarrashidi SA, Shekarchizadeh A (2015). "Multilevel Modeling of Science Achievement in the TIMSS Participating Countries." *The Journal of Educational Research*, **108**, 449–464. doi:10.1080/00220671.2014.917254.

Mullis IVS, Martin MO (eds.) (2013). *TIMSS 2015 Assessment Frameworks*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Mullis IVS, Martin MO (eds.) (2015). *PIRLS 2016 Assessment Framework, 2nd Edition*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Mullis IVS, Martin MO, Foy P, Drucker KT (2012). *PIRLS 2011 International Results in Reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Mullis IVS, Martin MO, Foy P, Hooper M (2016a). *TIMSS 2015 International Results in Mathematics*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Mullis IVS, Martin MO, Foy P, Hooper M (2016b). *TIMSS 2015 International Results in Science*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Mullis IVS, Martin MO, Foy P, Hooper M (2017). *PIRLS 2016 International Results in Reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill.

Muraki E (1992). "A Generalized Partial Credit Model: Application of an EM Algorithm." *Applied Psychological Measurement*, **16**, 159–176. doi:10.1177/014662169201600206.

Muthén B (1994). "Multilevel Covariance Structure Analysis." In J Hox, I Kreft (eds.), *Multilevel Modeling: A Special Issue of Sociological Methods & Research*, volume 22, pp. 376–398. Sage Publications.

Muthén LK, Muthén BO (2017). *Mplus User's Guide*. 8th edition. Muthén & Muthén, Los Angeles.

Myrberg E (2007). "The Effect of Formal Teacher Education on Reading Achievement of 3rd-Grade Students in Public and Independent Schools in Sweden." *Educational Studies*, **33**(2), 145–162. doi:10.1080/03055690601068311.

Natarajan R, Kass RE (2000). "Reference Bayesian Methods for Generalized Linear Mixed Models." *Journal of the American Statistical Association*, **95**(449), 227–337. doi:10.1080/01621459.2000.10473916.

Noble KG, McCandless BD, Farah MJ (2007). "Socioeconomic Gradients Predict Individual Differences in Neurocognitive Development." *Developmental Science*, **10**, 464–480. doi:10.1111/j.1467-7687.2007.00600.x.

Novikov L (2003). "A Remark on Efficient Simulations in SAS." *Journal of the Royal Statistical Society D*, **52**, 83–86. doi:10.1111/1467-9884.00343.

OECD (2009). *PISA Data Analysis Manual*. OECD Publishing, Paris.

OECD (2014). *PISA 2012 Technical Report*. OECD Publishing, Paris.

OECD (2016a). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD Publishing, Paris.

OECD (2016b). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*. OECD Publishing, Paris.

OECD (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving, Revised Edition*. OECD Publishing, Paris.

OECD (2017b). *PISA 2015 Technical Report*. OECD Publishing, Paris.

Pfeffermann D, Skinner CJ, Holmes DJ, Goldstein H, Rasbash J (1998). "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society B*, **60**(1), 23–40. doi:10.1111/1467-9868.00106.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2017). ***nlme**: Linear and Nonlinear Mixed Effects Models*. doi:10.32614/CRAN.package.nlme. R package version 3.1-131.

Pinheiro JC, Bates DM (1995). "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model." *Journal of Computational and Graphical Statistics*, **4**(1), 12–35. doi:10.2307/1390625.

Pinheiro JC, Chao EC (2006). "Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models." *Journal of Computational and Graphical Statistics*, **15**(1), 58–81. doi:10.1198/106186006x96962.

Pintrich PR, Smith DAF, Garcia T, McKeachie WJ (1991). *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning, Ann Arbor.

Rabe-Hesketh S, Skrondal A (2006). "Multilevel Modelling of Complex Survey Data." *Journal of the Royal Statistical Society A*, **169**(4), 805–827. doi:10.1111/j.1467-985x.2006.00426.x.

Rabe-Hesketh S, Skrondal A, Pickles A (2004). "Generalized Multilevel Structural Equation Modelling." *Psychometrika*, **69**(2), 167–190. doi:10.1007/bf02295939.

Rabe-Hesketh S, Skrondal A, Pickles A (2005). "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects." *Journal of Econometrics*, **128**(2), 301–323. doi:10.1016/j.jeconom.2004.08.017.

Rasch G (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The University of Chicago Press.

Raudenbush SW, Bryk AS (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, London.

Raudenbush SW, Bryk AS, Congdon R (2013). *HLM 7.01 for Windows*. Scientific Software International, Skokie.

Raudenbush SW, Yang ML, Yosef M (2000). "Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation." *Journal of Computational and Graphical Statistics*, **9**(1), 141–157. doi:10.2307/1390617.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rizopoulos D (2020). ***GLMMadaptive**: Generalized Linear Mixed Models Using Adaptive Gaussian Quadrature*. doi:10.32614/CRAN.package.glmmadaptive. R package version 0.6-8.

Rockwood N, Jeon M (2019). "Estimating Complex Measurement and Growth Models Using the R Package **PLmixed**." *Multivariate Behavioral Research*, **54**(2), 288–306. doi:10.1080/00273171.2018.1516541. R package version 0.1.2.

Rubin DR (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

Rust KF (1985). "Variance Estimation for Complex Estimators in Sample Surveys." *Journal of Official Statistics*, **1**, 381–397. doi:10.1177/096228029600500305.

Rust KF, Rao JNK (1996). "Variance Estimation for Complex Surveys Using Replication Techniques." *Statistical Methods in Medical Research: Special Issue on the Analysis of Complex Surveys*, **5**(3), 283–310. doi:10.1177/096228029600500305.

Rutkowski L, Gonzales E, Joncas M, Von Davier M (2010). "International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting." *Educational Researcher*, **39**(2), 142–151. doi:10.3102/0013189x10363170.

SAS Institute Inc (2016a). *SAS/IML 14.2 User's Guide*. Cary. URL http://support.sas.com/documentation/onlinedoc/iml/142/imlug.pdf.

SAS Institute Inc (2016b). *SAS/STAT 14.2 User's Guide*. Cary. URL http://support.sas.com/documentation/onlinedoc/stat/142/glimmix.pdf.

SAS Institute Inc (2017). *SAS/STAT 14.3 User's Guide*. Cary. URL http://support.sas.com/documentation/onlinedoc/stat/143/statug.pdf.

SAS Institute Inc (2020). *The SAS System, Version 15.2*. SAS Institute Inc., Cary. URL https://www.sas.com/.

SAS Institute Inc (2022). *SAS/STAT Software, Version 15.3*. SAS Institute Inc., Cary. URL http://www.sas.com/.

Schulz W, Ainley J, Fraillon J, Losito B, Agrusti G, Friedman T (2018). *Becoming Citizens in a Changing World: IEA International Civic and Citizenship Education Study 2016 International Report*. Springer-Verlag, Cham. doi:10.1007/978-3-319-73963-2.

Schulz-Heidorf K, Solheim OJ (2016). "Adapted Teaching: A Chance to Reduce the Effect of Social Origin? A Comparison Between Germany and Norway, Using PIRLS 2011." *Tertium Comparationis*, **22**(2), 230–259.

Schwarz G (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**(2), 461–464. doi:10.1214/aos/1176344136.

Shall R (1991). "Estimation in Generalized Linear Models with Random Effects." *Biometrika*, **78**(4), 719–727. doi:10.1093/biomet/78.4.719.

Shiner LB, Beal SL (1985). "Pharmacokinetic Parameter Estimates from Several Least Squares Procedures: Superiority of Extended Least Squares." *Journal of Pharmacokinetics and Biopharmaceutics*, **13**, 185–201. doi:10.1007/bf01059398.

Shun Z (1997). "Another Look at the Salamander Mating Data: A Modified Laplace Approximation Approach." *Journal of the American Statistical Association*, **92**(437), 341–349. doi:10.1080/01621459.1997.10473632.

Shun Z, McCullagh P (1995). "Laplace Approximation of High-Dimensional Integrals." *Journal of the Royal Statistical Society B*, **57**(4), 749–760. doi:10.1111/j.2517-6161.1995.tb02060.x.

Skrondal A, Rabe-Hesketh S (2004). *Generalized Latent Variable Modelling. Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL.

Smith AFM (1973). "A General Bayesian Linear Model." *Journal of the Royal Statistical Society B*, **35**(1), 67–75. doi:10.1111/j.2517-6161.1973.tb00937.x.

Smith DS, Wendt H, Kasper D (2016). "Social Reproduction and Sex in German Primary Schools." *Compare: A Journal of Comparative and International Education*. doi:10.1080/03057925.2016.1158643.

Snijders TAB, Bosker RJ (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd edition. Sage Publishers, London.

StataCorp (2019). *Stata Statistical Software: Release 16*. StataCorp LLC, College Station. URL https://www.stata.com/.

Sun L, Bradley KD, Akers K (2012). "A Multilevel Modelling Approach to Investigating Factors Impacting Science Achievement for Secondary School Students: PISA Hong Kong Sample." *International Journal of Science Education*, **34**(14), 2107–2125. doi:10.1080/09500693.2012.708063.

Tavsancil E, Yalcin S (2015). "A Determination of Turkish Students' Achievement Using Hierarchical Linear Models in Trends in International Mathematics-Science Study (TIMSS) 2011." *Anthropologist*, **22**(2), 390–396. doi:10.1080/09720073.2015.11891891.

Tuerlinckx F, Rijmen F, Verbeke G, De Boeck P (2006). "Statistical Inference in Generalized Linear Mixed Models: A Review." *British Journal of Mathematical and Statistical Psychology*, **59**(2), 225–255. doi:10.1348/000711005x79857.

Van De Pol J, Volman M, Oort F, Beishuizen J (2014). "Teacher Scaffolding in Small-Group Work: An Intervention Study." *Journal of the Learning Sciences*, **23**(4), 600–650. doi:10.1080/10508406.2013.805300.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Vock DM, Davidian M, Tsiatis AA (2014). "SNP_NLMM: A SAS Macro to Implement a Flexible Random Effects Density for Generalized Linear and Nonlinear Mixed Models." *Journal of Statistical Software, Code Snippets*, **56**(2), 1–21. doi:10.18637/jss.v056.c02.

Von Davier AA, Carstensen CH, Von Davier M (2008). "Linking Competencies in Horizontal, Vertical, and Longitudinal Settings and Measuring Growth." In H Hartig, E Klieme, D Leutner (eds.), *Assessment of Competencies in Educational Contexts*, pp. 121–149. Hogrefe & Huber Publisher, Göttingen.

Von Davier M, Gonzalez E, Mislevy RJ (2009). "What Are Plausible Values and Why Are They Useful?" *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, **2**, 9–36. doi:10.1007/978-3-030-47515-4_3.

Von Davier M, Sinharay S, Oranje A, Beaton A (2007). "The Statistical Procedures Used in National Assessment of Educational Progress: Recent Developments and Future Directions." In CR Rao, S Sinharay (eds.), *Handbook of Statistics, Vol. 26*, chapter 32, pp. 1039–1055. Elsevier, Amsterdam.

Webster BJ, Fisher DL (2000). "Accounting for Variation in Science and Mathematics Achievement: A Multilevel Analysis of Australian Data Third International Mathematics and Science Study (TIMSS)." *School Effectiveness and School Improvement*, **11**(3), 339–360. doi:10.1076/0924-3453(200009)11:3;1-g;ft339.

Wendt H, Kasper D, Trendtel M (2017). "Assuming Measurement Invariance of Background Indicators in International Comparative Educational Achievement Studies: A Challenge for the Interpretation of Achievement Differences." *Large-Scale Assessment in Education*, **5**(10), 1–34. doi:10.1186/s40536-017-0043-9.

White H (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, **50**(1), 1–25. doi:10.2307/1912526.

Wiberg M, Rolfsman E (2013). "School Effectiveness in Science in Sweden and Norway Viewed from a TIMSS Perspective." *Utbildning & Demokrati*, **22**(3), 69–84. doi:10.48059/uod.v22i3.1003.

Wolfinger RD, O'Connell MA (1993). "Generalized Linear Mixed Models: A Pseudo-Likelihood Approach." *Journal of Statistical Computation and Simulation*, **48**(3–4), 233–243. doi:10.1080/00949659308811554.

Wolter KM (2007). *Introduction to Variance Estimation*. Springer-Verlag, New York.

Woltman H, Feldstain A, MacKay JC, Rocchi M (2012). "An Introduction to Hierarchical Linear Modeling." *Tutorials in Quantitative Methods for Psychology*, **8**(1), 52–69. doi:10.20982/tqmp.08.1.p052.

Wright BD, Masters GN (1982). *Rating Scale Analysis*. MESA Press, Chicago.

Zeger SL, Karim MR (1991). "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach." *Journal of the American Statistical Association*, **86**(413), 79–86. doi:10.1080/01621459.1991.10475006.

Zhu M (2014). "Analyzing Multilevel Models with the GLIMMIX Procedure." In *Proceedings of the SAS Global Forum 2014 Conference*. SAS Institute Inc, Cary. Paper SAS026-2014, URL http://support.sas.com/resources/papers/proceedings14/SAS026-2014.pdf.

A. The SURVEYHLM macro code

In the next two sections, we set out the **SURVEYHLM** code. Section A.1 presents the code for calculating means, variances, and frequencies, while Section A.2 provides the code for calculating multilevel models.

A.1. Means, variances, and frequencies

The **SURVEYHLM** macro produces, by default, descriptive statistics for all variables in the multilevel model. The statistics for each continuous variable include the following: the total number of observations, the number of missing observations, the number of non-missing observations, the minimum and the maximum values, and the arithmetic mean, standard error, variance, and standard deviation. The statistics for categorical variables include the frequencies per category, the cumulative frequencies, the percentages, the standard errors of these percentages, and the cumulative percentage. If the analysis includes use of a combined weight (i.e., `&wgt` exists), the macro weights the results, and if it requests use of a repeated replication technique (i.e., either `&jkrep` and `&jkzone` or `&rwnames`), the standard errors are based on this technique. In addition, if the analysis uses plausible values as the dependent variable (i.e., `&npv > 1`) and requests use of the repeated replication technique, the macro uses either all plausible values to calculate the standard errors (i.e., `&shrtcut = n`) or just the first plausible value (i.e., `&shrtcut = y`).

Before we present the syntax for creating these descriptive statistics, please note in particular the syntax for the macro parameter `&dataset`. This parameter makes it possible to distinguish different datasets because it combines the technique that needs to be used to calculate standard errors and the available weights (see Table 2).[13] Assume, for example, that the user has supplied replication weights. If the macro parameters `&l2wgt`, `&l1wgt`, and `&wgt` also exist, then `&dataset = 1`. If, however, no weight is specified, then `&dataset = 4`. The combination of these different characteristics thus distinguishes 18 datasets. However, it should be mentioned that the macro variable `&dataset` is an internal variable, which means users must not define it. Instead, `&dataset` automatically defines the variable, with the definition based on the information supplied during invocation of the **SURVEYHLM** macro.

	Existing weighting parameter					
SE	L2/L1/WGT	L2/L1	WGT	None	L3/L2/L1/WGT	L3/L2/L1
REPW	1	2	3	4	13	14
JACK	5	6	7	8	15	16
MODL	9	10	11	12	17	18

Table 2: Datasets differentiated via `&dataset`.

[13] Table 2 uses abbreviations for the names of the macro parameter. For example, `L2` means level 2 weights (i.e., `&l2wgt` exists). The name of the technique used to calculate the standard errors (the column labeled SE) is also abbreviated. REPW means that calculation of the standard errors is based on the repeated replication technique and includes user-supplied replication weights. JACK indicates that the jackknife technique is being used for calculating the standard errors, and MODL denotes sandwich estimators for the standard errors.

Means and variances

We will first present the syntax for calculating the means and variances when `&npv = 1` and then the syntax for `&npv > 1`.

```
proc iml;
use &datnr;
read all var {&dvnames &xvar1 &xvar2 &xvar3} into y3;

%if &dataset. in 2|4|6|8|10|12|14|16|18 %then %do;
 wgt = j(nrow(y3), 1, 1);
%end;
%else %if &dataset. in 1|3|5|7|9|11|13|15|17 %then %do;
 use &datnr;
 read all var {&wgt} into y0;
 wgt = y0;
%end;
 dvw = y3 # wgt;
 wcs = dvw[+, ];
 wgtmis = wgt # (y3 ^= .);
 sumwgt = wgtmis[+, ];
 m = wcs / sumwgt;
 mt = t(m);
 dif = y3 - m;
 dif2dif = dif # dif;
 difw = dif2dif # wgt;
 difs = difw[+, ];
 var = difs / (sumwgt - 1);
 vart = t(var);
 std = sqrt(var);
 stdt = t(std);
```

In the first part of the syntax, the variables whose names appear in `&dvnames` (the continuous dependent variables), `&xvar1` (the level-1 continuous predictors), `&xvar2` (the level-2 continuous predictors), or `&xvar3` (the level-3 continuous predictors) are written to the vector `y3`. The second part of the syntax estimates the means and the variances for the `y3` variables. If the analysis is unweighted, a vector of ones for `wgt` is created, otherwise the weights are used. Before the remaining calculations are carried out, the `wgt` vector weights the individual values to obtain `dvw`. Thus, the mean estimates for the variables `m` are based on the sum of weighted individual values across observations `wcs` divided by the sum of weights across the non-missing observations `sumwgt`. In a step that corresponds with the weighting of the individual values, the squared differences between the individual values and the means are also weighted to obtain `difw`. The sum of these squares per variable (i.e., `difs`) is divided by the degrees of freedom `sumwgt - 1` to obtain the row vector `var`, with the variances for each variable in the columns.

If the user requests standard errors for the means, these will simply be the standard deviations (i.e., the square root of `var`) of the variables divided by the square root of `n` (see the following code). However, if the user wants to calculate standard errors, with these ideally based on

the repeated replication technique, a few more steps are required. First, the variables whose names appear in `&dvnames`, `&xvar1`, `&xvar2`, or `&xvar3` and that are in the dataset with the name `&datndesc0`[14] are read into the matrix `p1`. The first row of `p1` is then written into the vector `d1`, and the remaining rows are written into the matrix `d2`. The squared differences of each row of `d2` from `d` are calculated in the do loop, and these differences combine to form the matrix `pdifall`. The sum of these squared differences per variable (i.e., `pdifall[+, ]`) is then multiplied by `&jkfac` (usually `&jkfac = 0.5` or `&jkfac = 1`, see Section 2) to obtain `totvar`, and the square root of `totvar` gives the replication variance for the variables.

```
%if &dataset. in 9|10|11|12|17|18 %then %do;
 n = countn(y3, "col");
 stder = std / sqrt(n);
%end;
%else %if &dataset. in 1|2|3|4|5|6|7|8|13|14|15|16 %then %do;
 use &datndesc0;
 read all var{&dvnames &xvar1 &xvar2 &xvar3} into p1;
 d = p1;
 d1 = p1[1, ];
 d2 = p1[2:&njk1, ];
 do i = 1 to &nrwgt;
  pdif = (d1 - d2[i, ]) ## 2;
  pdifall = pdifall // pdif;
 end;
 jkvar = pdifall[+, ];
 totvar = &jkfac * jkvar;
 stder = totvar ## .5;
%end;
stdet = t(stder);
Out2 = mt || stdet || vart || stdt;
```

The syntax for calculating the descriptive statistics when `&npv > 1` strongly resembles the syntax for the `&npv = 1` case. In practical terms, differences occur only for the dependent variables. If plausible values are used as the dependent variable, then the mean, standard deviation, and standard errors for the dependent variable must heed the measurement error of this variable. The mean is now the arithmetic mean across all `&npv` plausible values, while the standard deviation and the standard error take into account the imputation variance (i.e., the variance between the plausible values). As can be seen below, the estimate of the mean `gm` is now the arithmetic mean of the `&npv` means (one for each plausible value), and the imputation variance `impvar` is a function of the variance between the `&npv` means of the dependent variable. The standard deviation is calculated by averaging the `&npv` variances (one for each plausible value) and adding the imputation variance to this term.

```
outs = out2[1:&npv, 1];
gm = outs[:, ];
impv = var(outs);
```

[14] The dataset `&datndesc0` contains the means for each replicated sample `i` and the means for the total sample. Thus, the first row of `&datndesc0` contains the sample means for the total sample, the second row the sample means for the first replicated sample, and so on.

```
impvar = %sysevalf(%sysevalf(&npv + 1) / &npv) * impv;
vars = var[, 1:&npv];
mvar = (vars[, +]) / (&npv);
mvars = mvar + impvar;
std = mvars ## .5;
%if &dataset. in 9|10|11|12|17|18 %then %do;
 stder = std / sqrt(n[1, 1]);
%end;
%else %if &dataset. in 1|2|3|4|5|6|7|8|13|14|15|16 %then %do;
 totvars = totvar[, 1:&npv];
 %if %lowcase(&shrtcut) = y %then %do;
  totvar = totvars[, 1] + impvar;
  stder = totvar ## .5;
 %end;
 %else %if %lowcase(&shrtcut) = n %then %do;
  totvar = (totvars[, +]) / (&npv);
  totvar1 = totvar + impvar;
  stder = totvar1 ## .5;
 %end;
%end;
```

The procedure for calculating the standard error of the mean depends on the macro parameter `&shrtcut`. For `%lowcase(&shrtcut) = y`, only the replication variance of the first plausible value is used, and the imputation variance must be added to the replication variance. Thus, instead of using `totvar = &jkfac * jkvar`, we use `totvar = (&jkfac * jkvar) + impvar`. When `%lowcase(&shrtcut) = n`, the `&npv` replication variances are averaged, and the imputation variance is added to this average value. Thus, in comparison to the code above, `totvar = k + impvar` with `k = jkvar1[, +] / &npv` and `jkvar1 = &jkfac * jkvar`.

Frequencies

The `PROC FREQ` function of SAS is used, within a do loop, to calculate the frequencies for each variable `&cvar` (i.e., for each dependent or independent categorical variable of the GLMM model). The estimated statistics in the data steps following the frequency function are rounded, and the standard error for the percentage per category is calculated. In addition, the `retain` statement determines the order in which the statistics will be presented in the eventual printed-out dataset.

```
%do t = 1 %to &ncvar;
 %let cvarf = %scan(&cvar, &t);
 proc freq data = &datnr;
 tables &cvarf / outcum out = &datndesc3._&t.;
 &wgtn;
 run;
%end;

%do u = 1 %to &ncall;
 %let cvarf = %scan(&cvarp, &u);
 data svyhlm_&cvarf._print (keep = &cvarf Frequency Percent CumFrequency
```

```
 CumPercent StdErr);
 set &datndesc3._&u.;
 if percent = . then delete;
 countn = round(count, 1);
 percentn = round(percent, 0.01);
 cum_freqn = round(cum_freq, 1);
 cum_pctn = round(cum_pct, 0.01);
 if stderr = . then do;
  stderr1 = round(sqrt(percent * (100 - percent) / &nobs), 0.001);
  stderr = round(stderr1, 0.01);
  drop stderr1;
 end;
 drop count percent cum_freq cum_Pct;
 rename countn = Frequency percentn = Percent cum_freqn = CumFrequency
 cum_pctn = CumPercent;
 run;
 data svyhlm_&cvarf._print;
 retain &cvarf Frequency CumFrequency Percent StdErr CumPercent;
 set svyhlm_&cvarf._print;
 run;
%end;
```

If standard errors need to be based on the repeated replication technique, then the frequency procedure is applied `&nrwgt` times to each variable `&cvar` with the appropriate weighting variable `rwgt&k` within a do loop (not depicted here). After the variables have been rounded and reordered, the `&nrwgt` datasets with the replicated estimates of the statistics are combined into the dataset `&datndesc4`, and the replication-based estimates of the standard errors are calculated again, this time with the SAS's `PROC IML` procedure. In essence, the `IML` syntax used to calculate the standard error of the mean (see above) can also be used here.

With respect to `&npv > 1`, the frequencies for each plausible value are calculated within a do loop (not depicted here), the results of the frequency procedures are then averaged across the different plausible values, and the dataset is readied for presentation in printed form. If a user requests that replication variance be generated for the percentages of each category of the dependent variable, further analysis, involving three steps, is necessary. During the first step, the `PROC MIANALYZE` procedure of SAS calculates the between imputation variance for the percentages. The second step involves estimation of the sample statistics for the replication samples (not depicted here), and the third involves calculation of the replication variance. Once again, the syntax used is basically the same as the `IML` syntax presented above. However, as can be seen from the next set of syntax, users need to take account of the imputation variance, that is the variance of `percent` across the `&npv` plausible values.[15]

```
proc iml;
use svyhlm_&cvarf._print;
read all var {Percent} into y1;
%if &ntvar > 1 and &t > &ncvars %then %do;
```

[15] `svyhlm_&cvarf._print` is a dataset with the percentages of the whole sample, whereas `svyhlm_pcfreq_vi_m` is a dataset with the results of the `PROC MIANALYZE` procedure.

```
 use svyhlm_pcfreq_vi_m;
 read all var {BetVar} into y3;
 impvar = %sysevalf(%sysevalf(&npv + 1) / &npv) * y3;
%end;
```

Because calculation of repeated replication-based standard errors when `%lowcase(&shrtcut) = n` uses essentially the same procedure as the one described for `%lowcase(&shrtcut) = y`, we decided not to present its code here. However, it is important to remember that the aforementioned procedure must be applied separately to each plausible value, and the `&npv` different estimates for `totvar` must be averaged so that the square root of this average becomes the repeated replication-based standard error.

A.2. Multilevel analysis

The syntax code for multilevel analyses is, not surprisingly, more detailed than that of the code presented in Section A.1. Here, the syntax code is used to prepare the dataset (e.g., center the variables, scale the weights, construct level-specific replication weights) and define the model components (e.g., the number of iterations, the estimation method, the type of covariance structure for the random components) and/or the post-processing of the results (e.g., checking for convergence of the estimation, transforming the covariance matrix of the random components into a correlation matrix, and preparing the results for print presentation). However, because discussion of these details of the syntax is not feasible within the scope of this article, we focus only on the code's essential elements.

Estimating the multilevel model

We used the `PROC GLIMMIX` procedure from SAS to estimate the multilevel model.[16] The multilevel analyses are performed in an inner loop (for the repeated replication estimates) and an outer loop (for the different plausible values).[17] We then used the dataset defined

[16] Ene, Leighton, Blue, and Bell (2015) and Zhu (2014) have shown that `PROC GLIMMIX` can be used for multilevel analyses in SAS by applying this procedure, among others, to the same Programme for International Student Assessment (PISA) data introduced by Rabe-Hesketh and Skrondal (2006). We therefore decided to use that procedure in our **SURVEYHLM** macro. We could have used, as an alternative, `PROC NLMIXED`, which is what Anderson, Kim, and Keller (2014), Kurada (2016), and Vock, Davidian, and Tsiatis (2014) did. However, if level-specific weights need to be used, then the user needs to specify the `replicate` statement (Anderson *et al.* 2014). Unfortunately, if more than one `random` statement is used, which is necessary, for example, for a three-level model, then the statement `replicate` is not allowed (SAS Institute Inc. 2016b). Hence, `PROC NLMIXED` cannot be used for multilevel analyses with level-specific weights and more than two levels. Another way to estimate multilevel models in SAS is through use of the `PROC IML` procedure (SAS Institute Inc. 2016a). However, it implies, among other considerations, that the objective function of the multilevel model and the optimization procedure must be specified manually. In addition, all desirable statistics that are additional to the multilevel coefficients and the random components and are part of the objective function must be calculated separately. Such statistics include Akaike's information criteria (Akaike 1974) and Schwarz's Bayesian criterion (Schwarz 1978). Because `PROC GLIMMIX` provides these additional statistics without additional code, the use of `PROC IML` seems inefficient.

[17] In preliminary versions of our macro, we used the `BY` processing features of SAS instead of the loops presented here. However, our results suggest that the running time of the macro will decrease when the `BY` processing feature is used. Hence, we decided to implement the loops in this final version. We nevertheless strongly recommend use of the `BY` processing features of SAS whenever possible (Cassell 2007; Novikov 2003). In general, using the `BY` processing future of SAS is more efficient, is less error-prone, and needs less calculation time than looping.

by name in the macro parameter `&datnrlm` to invoke the `GLIMMIX` procedure. Although this dataset was essentially the same as the dataset the user supplied for the **SURVEYHLM** model, it now included, among other features and, if requested, the scaled weights and the replication weights. The macro parameter `&method` specifies the estimation method, and if the model has no random effects, the method is `&method = method = mspl`, which means that the estimation method for a model such as this one is the maximum pseudo-likelihood. For all other models, `&method = method = quad(&qpoints)`. Consequently, the estimation methods used for these models are the maximum likelihood estimation with Gauss-Hermit quadrature and, if specified, the user-specified number of quadrature points `&qpoints`. We also used the option `empirical = classical`, which implies that the standard errors for the multilevel coefficients will be based on the sandwich variance estimator whenever the user does not specify the repeated replication-based technique.

```
%do pv = 1 %to &npv;
 %do i = 0 %to &nrwgtc;
  proc glimmix data = &datnrlm &method &plotsm empirical = classical;
  nloptions gconv = &gconv &maxf &maxi technique = &tec;
  class &sub3 &sub2 &cvar1n &cvar2n &cvar3n;
  model &y = &xvar1n &xvar2n &xvar3n &cvar1n &cvar2n &cvar3n / &nointr
  solution dist = &dist &obsweightc;
  %if &mod. in 1|2|3|4 %then %do;
   random &inter2 &res &rslopen / sub = &sub2 g &gs &rweight2c type = &type2
   &ldatan2;
  %end;
  %else %if &mod. in 5|6|7|8|9|10|11|12|13 %then %do;
   random &inter2 &rslopen / sub = &sub2(&sub3) g &gs &rweight2c
   type = &type2 &ldatan2;
   random &inter3 &rslopen3 / sub = &sub3 g &gs &rweight3c type = &type3
   &ldatan3;
  %end;
  ;
  %if &i = 0 and &pv > 1 and %lowcase(&start) = y %then %do;
   parms / pdata = svyhlm_vp_1;
  %end;
  ;
  &wgtlmc ;
  ods output CovParms = svyhlm_vp_&pv._jk_&i.
  ParameterEstimates = svyhlm_p_&pv._jk_&i. &convst
  &gmat &constat &gv &fitstat &nobsglm ;
  run;
 %end;
%end;
```

Note that in the row beginning with the statement `nloptions`, we define some options for the nonlinear optimization procedure. We specify the relative gradient convergent criterion with `gconv` and `&gconv = 1E-8` as the default. The macro parameters `&maxf` and `&maxi` define the maximum number of function calls and the maximum number of iterations during the

&mod	Levels	FIX	RI2	RS2	RI3	RS3
1	2	✓	✓			
2	2	✓	✓	✓		
3	1	✓				
4	2	✓		✓		
5	3	✓	✓		✓	
6	3	✓	✓			✓
7	3	✓	✓		✓	✓
8	3	✓		✓	✓	
9	3	✓		✓		✓
10	3	✓		✓	✓	✓
11	3	✓	✓	✓	✓	
12	3	✓	✓	✓		✓
13	3	✓	✓	✓	✓	✓

Table 3: Possible values for &mod and models that can be fitted with the **SURVEYHLM** macro.

optimization. The default values for this parameter depend on the optimization technique (i.e., on &tec) that is used in the technique statement. (For details of possible options for &gconv, &maxf, &maxi, and &tec, see the documentation for the nloptions statement for the PROC GLIMMIX procedure in SAS Institute Inc. 2017.)

The class statement defines the variables that should be considered as categorical. These are the level identifiers &sub3 and &sub2 and the categorical predictor variables &cvar1n, &cvar2n, and &cvar3n. The identifier for the fixed effects model is &sub2 = &nest1, else &sub2 = &nest2; for the three-level models it is &sub3 = &nest3, and the model statement defines the dependent variable with the name &y. Also included as independent variables are the continuous predictor variables &xvar1n, &xvar2n, and &xvar3n, and the categorical predictor variables &cvar1n, &cvar2n, and &cvar3n. Several options are specified after the front slash. The macro parameter &nointr defines whether an intercept should be included in the model, and because &nointr is per default empty, an intercept is included. If an intercept is not included in the model, then the user must specify noint = y when invoking the %SUREVYHLM macro. The statement solution requests a solution for the fixed effects parameters, and dist specifies the probability distribution of the data. (For details about possible options for &dist, see the documentation for the PROC GLIMMIX procedure in SAS Institute Inc. 2017.) The default is &dist = normal, which means a normally distributed dependent variable is assumed. If the user requests a level-1 specific weight, this appears in &obsweightc.

The following statement (i.e., random) is conditional, which means it is executed only when &mod = 1, &mod = 2, &mod = 3, or &mod = 4. The macro parameter &mod defines the model that the user requests.[18] Table 3 presents an overview of the possible values for &mod. Thus, for example, when &mod = 1, the estimated model will be a two-level model with a random intercept on level 2, and when &mod = 5, the requested model will be a three-level one with

[18]The macro variable &mod is an internal variable, which means it is based on the information the user supplied during invocation of the **SURVEYHLM** macro. It also means that appropriate values to &mod will be assigned automatically.

a random intercept on level 2 and random slopes on level 3. The conditional statement of `random` therefore implies that this statement will only be executed if the analysis involves a fixed effects model or a level-2 model.

The macro parameter `&inter2` in the `random` statement is `&inter2 = intercept` if a random intercept is assumed, else `&inter2` is empty. The `&res` parameter is defined only for the fixed effects model, that is, when `&mod = 3`. In that case, `&res = _residual_`, meaning that the residual on level 1 is assumed to be random. If the user defines the random slopes on level 2 (i.e., if `&rslope2` is not empty), then the names of these variables occur in `&rslopen`. Several options are available after the front slash of the `random` statement. The option `sub` defines the subject in the specified model, and complete independence is assumed across the subjects. If the model is a fixed effects one, `&sub2 = &nest1`, the subject across which the `&res = _residual_` varies randomly is the level-1 identifier. For all other models, `&sub2 = &nest2`, which means the random effects specified in `&inter2` and `&rslopen` vary randomly across the level-2 identifier. The `G` option displays the estimated random components of the $\boldsymbol{G}$ matrix, and `&rweight2c = weight = &l2wgts` defines the level-2 specific weight, if the user supplies this. The `type` argument specifies the structure of $\boldsymbol{R}$ when `&mod = 3` and the structure of $\boldsymbol{G}$ when `&mod ~= 3`. The default type is `&type2 = vc`, which means a distinct variance component has been assigned to each random effect and the covariances between the random effects are assumed to be zero. (For other possible options of the type argument, see SAS Institute Inc. 2016b.) The macro parameter `&ldatan2` is necessary when `&type2 = lin(q)`. In this instance, `&ldatan2 = ldata = &ldata2`, and `&ldata2` defines the dataset with the matrices of the assumed linear combination.

If the user analyzes a three-level model, two `random` statements are executed. The first is for the level-2 random components, and the second is for the level-3 random components. Although the meaning of the macro parameter is nearly the same as that for the two-level model, the `&rweight`, `&type`, and `&ldatan` arguments are now available for both levels. In addition, the first `sub` statement shows that if the model is a three-level one, we can assume that the level-2 units (i.e., `&sub2`) are nested within the level-3 unit (i.e., `&sub3`).

The `&wgtlmc` parameter is defined only when `&dataset = 3`, `&dataset = 7`, or `&dataset = 11`. In this case, `&wgtlmc = weight &wgt`, and `&rweight2c` and `&rweight3c` will be empty. The `ods output` statement defines the different output datasets. Fit statistics, number of observations, random components together with their standard errors, and the fixed effects parameter estimates are written to the working directory for all possible models. Information about the convergence status, conditional fit statistics, and the $\boldsymbol{G}$ matrix is also written to the working directory for all models, except the fixed effects model, and all datasets are used afterwards for generating the print output. The information about the convergent is also used to determine if the analysis will continue. If convergence using the whole sample does not occur, the analysis will be interrupted.

Estimating repeated replication-based standard errors

The procedure for estimating the repeated replication-based standard errors for the fixed effects and the random components consists of two steps. First, the appropriate `PROC GLIMMIX` syntax is run `&nrwgt` times, during which each pass is made with the corresponding replication weight (see the previous section). Second, the replication-based variance is estimated within a `PROC IML` step. We do, however, need to comment on the convergent information. If

a replication-based PROC GLIMMIX procedure does not converge, the fixed effects parameter estimates and the random component estimates from the appropriate full model will be used as the default output for this run. However, a message will be sent to the SAS log that tells the user the number of &nrwgt replications that did not converge.

The fixed effects parameter estimates and the random component estimates of the &nrwgt replication-based analysis are each combined into a single dataset, and these datasets are used in the PROC IML step that calculates the replication-based standard errors. As can be seen from the code for &npv > 1 and %lowcase(&shrtcut) = n set out below, the replication variance for each of the &npv plausible values is calculated in a do loop, written into the object E, and outputted to the dataset surveyhlm_p_iml1_&pv.

```
%do pv = 1 %to &npvr;
 %let tempjkp = %NewDatasetName(temp);
 data &tempjkp;
 merge %namep(svyhlm_p_&pv._jk_, &nrwgt);
 run;
 %let tempjkv = %NewDatasetName(temp);
 data &tempjkv;
 merge %namep(svyhlm_vp_&pv._jk_, &nrwgt);
 run;
 proc iml;
 start e(p, p1, p2) global(E);
 pp = p[, 1];
 do i = 1 to &nrwgt;
  pdif = (pp - p1[, i]) ## 2;
  pdifall = pdifall || pdif;
 end;
 jkvar1 = pdifall[, +];
 jkvar = &jkfac * jkvar1;
 %if &npv > 1 %then %do;
  impv = p2[, 1];
  impvar = %sysevalf(%sysevalf(&npv + 1) / &npv) * impv;
  totvar = jkvar + impvar;
 %end;
 %else %do;
  totvar = jkvar;
 %end;
 E = totvar;
 finish e;
 use svyhlm_p_&pv.;
 read all var{Estimate} into p;
 use &tempjkp;
 read all var _num_ into p1;
 %if &npv > 1 %then %do;
  use svyhlm_p_vi_m;
  read all var{BetVar} into p2;
 %end;
```

```
 %else %do;
  use svyhlm_p_&pv.;
  read all var{Estimate} into p2;
 %end;
 run e(p, p1, p2);
 create svyhlm_p_iml1_&pv. from E[colname = {"Tot&pv."}];
 append from E;
 close svyhlm_p_iml1_&pv.;
%end;
```

The datasets `surveyhlm_p_iml1_&pv.` are then combined into a single dataset, and `PROC IML` is used to average the `&npv` replication variance estimates. The square root of this average is the desired estimate of the replication-based standard errors (one for each fixed effect).

```
%let tempjkpc = %NewDatasetName(temp);
data &tempjkpc;
merge %namep(svyhlm_p_iml1_, &npvr);
run;
proc iml;
start f(p, p1) global(F);
pf = p[, 1];
df = p[, 2];
totvar = p1;
totvar1 = (totvar[, +]) / (&npvr);
stder = totvar1 ## .5;
tvalue = divide(pf, stder);
do i = 1 to nrow(df);
 if df[i] > 0 then prtc = 1 - probt(abs(tvalue[i]), df[i]);
 else prtc = .;
 prt = prt // prtc;
end;
F = stder || df || tvalue || prt;
finish f;
use svyhlm_p_print;
read all var{Estimate DF} into p;
use &tempjkpc;
read all var _num_ into p1;
run f(p, p1);
%let tempjkpr = %NewDatasetName(temp);
create &tempjkpr from F[colname = {"StdErr" "DF" "tvalue" "Probt"}];
append from F;
close &tempjkpr;
quit;
```

Affiliation:

Daniel Kasper
Department of General, Intercultural and International Comparative Education, and Educational Psychology
Faculty of Education
UHH Universität Hamburg
20146 Hamburg, Germany
E-mail: daniel.kasper@uni-hamburg.de
URL: https://www.ew.uni-hamburg.de/ueber-die-fakultaet/personen/kasper-d.html

Journal of Statistical Software https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
July 2025, Volume 113, Issue 5 *Submitted:* 2019-11-22
doi:10.18637/jss.v113.i05 *Accepted:* 2023-09-18