


METHODOLOGY

Open Access



Sampling weights in multilevel modelling: an investigation using PISA sampling structures

Julia Mang^{1*} , Helmut Küchenhoff², Sabine Meinck³ and Manfred Prenzel⁴

*Correspondence:

Julia.Mang@tum.de

¹ TUM School of Education,
Centre for International
Student Assessment
(ZIB), Technical University
of Munich (TUM), Arcisstr. 21,
80333 Munich, Germany
Full list of author information
is available at the end of the
article

Abstract

Background: Standard methods for analysing data from large-scale assessments (LSA) cannot merely be adopted if hierarchical (or multilevel) regression modelling should be applied. Currently various approaches exist; they all follow generally a design-based model of estimation using the pseudo maximum likelihood method and adjusted weights for the corresponding hierarchies. Specifically, several different approaches to using and scaling sampling weights in hierarchical models are promoted, yet no study has compared them to provide evidence of which method performs best and therefore should be preferred. Furthermore, different software programs implement different estimation algorithms, leading to different results.

Objective and method: In this study, we determine based on a simulation, the estimation procedure showing the smallest distortion to the actual population features. We consider different estimation, optimization and acceleration methods, and different approaches on using sampling weights. Three scenarios have been simulated using the statistical program R. The analyses have been performed with two software packages for hierarchical modelling of LSA data, namely Mplus and SAS.

Results and conclusions: The simulation results revealed three weighting approaches performing best in retrieving the true population parameters. One of them implies using only level two weights (here: final school weights) and is because of its simple implementation the most favourable one. This finding should provide a clear recommendation to researchers for using weights in multilevel modelling (MLM) when analysing LSA data, or data with a similar structure. Further, we found only little differences in the performance and default settings of the software programs used, with the software package Mplus providing slightly more precise estimates. Different algorithm starting settings or different accelerating methods for optimization could cause these distinctions. However, it should be emphasized that with the recommended weighting approach, both software packages perform equally well. Finally, two scaling techniques for student weights have been investigated. They provide both nearly identical results. We use data from the Programme for International Student Assessment (PISA) 2015 to illustrate the practical importance and relevance of weighting in analysing large-scale assessment data with hierarchical models.

Keywords: Sampling weights, Hierarchical models (HLM), Multilevel models (MLM), Programme for International Student Assessment (PISA), Large-scale assessment (LSA), Scaling of sampling weights

Introduction and theoretical framework

As is widely known in the field of large-scale assessments (LSAs), conducting a census survey is not productive from an organisational, time and most of all financial perspective (Rutkowski et al., 2010). Therefore, for many LSAs a two-stage stratified cluster sampling procedure is applied. More specifically, schools are sampled in a first step, in most cases using probability proportional to size (PPS) mechanism with stratification, i.e., larger schools are sampled with higher probability (Brewer & Hanif, 1983). In a second step, students are selected randomly within these sampled schools (OECD, 2017).

The aim of LSAs is to draw conclusions for a whole population by means of the chosen sample. For analysing those student samples, special weights for all sampling units (e.g., schools, classes, and students) are provided in order to avoid bias due to these sampling techniques (Meinck, 2020; OECD, 2017). Those weights reflect the selection probabilities of the schools and students, adjusted for non-response, and thereby the proportion of the population represented by each sampled school and student. The “Methods” section of this paper elaborates exemplarily the sampling procedure of the Programme for International Student Assessment (PISA), illustrated by an exemplary country (Germany).

As students within one school often are more similar to each other than students attending different schools, considering a hierarchical (or “multilevel”) model in analysing students is advisable. This is because such models better reflect the true multilevel structure of the education system with pupils nested within classes, schools and school systems. Furthermore, the cluster effects on sampling errors are taken into account in such models, which otherwise have to be reflected by using special complex estimation procedures [e.g., balanced repeated replication in PISA; OECD (2017)].

Even though the typical hierarchical structure in education includes three or even more levels (e.g., students within classes within schools within countries etc.), this article focuses on two levels, with students at level one and schools at level two. This is for several reasons. First, the general sampling scheme of several LSA such as PISA or the International Computer and Information Literacy Study (ICILS; Gebhardt et al., 2014) do not include class sampling at all. Second, if class sampling is incorporated, the actual (true) or sampled number of classes within schools is always small: often just one or two classes are sampled, especially in small schools. Therefore, it is impossible to disentangle class and school effects. Finally, research applying three-level models is sparse, probably because (i) not many datasets fulfil the necessary preconditions, (ii) models often do not converge, and (iii) because interpretation becomes more complex when adding levels and can be very challenging. Instead, most cross-national research with multilevel models uses also two-level models: identical models are run separately for each educational system participating in a specific assessment and are then compared. Hence, this contribution is fully valid for cross-country analysis.

Although there is sufficient evidence that sampling weights must be used in multilevel modelling (MLM) to obtain unbiased estimates (e.g. Cai, 2013), and also on how these weights should be used in single-level analyses, there is little discussion in the literature about which and how to use sampling weights in MLM. Asparouhov (2006) claims that data sets from studies with complex sampling designs are made available with weights prepared for, e.g., computing means, but that these weights are not appropriate for

multilevel models and can produce erroneous results if used in hierarchical analyses. Stapleton (2002) addresses the use of different weighting techniques. Rutkowski et al. (2010) argue that issues of weight scaling and parameter estimation are important considerations. They suggest a procedure for manually calculating appropriate weights at the levels of interest for analysis, using the design weights and nonresponse adjustments at each sampling stage for composing these level-specific weights. Carle (2009) recommends to rely on scaled weighted estimates rather than unscaled weighted ones.

Currently, four different approaches on how to use sampling weights in hierarchical models are recommended by different authors. Partly, different approaches are even recommended and used for the same type of data, leaving scholars in dubiety, which approach to use. They mainly relate to specific LSA, namely PISA, the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS; Martin & Mullis, 2013), the International Civic and Citizenship Education Study (ICCS) (Schulz et al., 2018) and ICILS (Gebhardt et al., 2014). The simulation study scenarios are based on these approaches, hence, a detailed description can be found in section “[Analysis procedures](#)”. In the following, we explain the technical background on how these weights can be scaled and incorporated for parameter estimation.

Pfeffermann et al. (1998) and Asparouhov (2006) advise to use a pseudo maximum likelihood approach for calculating estimates within and between the different levels using probability weighted generalized least squares (PWGLS) maximisation technique in order to obtain unbiased estimates. Alternatively, Rabe-Hesketh and Skrondal (2006) provide the expectation–maximisation techniques for maximizing the pseudo likelihood. No previous research includes a straightforward suggestion on how to scale level one weights in order to account for hierarchical structures. Three different approaches have been discussed in the literature (Graubard & Korn, 1996; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006) whereas only two approaches are applicable for survey data.

Several simulation studies (Asparouhov, 2004; Bertolet, 2008; Korn & Graubard, 2003; Rabe-Hesketh & Skrondal, 2006) conclude that there is no estimation procedure or adjustment of the weighting to be clearly preferred. Rather, the sampling design itself is decisive for the choice of the estimation procedure. Furthermore, different software programs implement different inference estimation methods, leading to different results (Chantala & Suchidnran, 2006; Chantala et al., 2011; West & Galecki, 2012).

Nevertheless, none of the papers so far has provided a comprehensive overview of all possible and previously used weighting approaches, a research gap that will be filled with this study. The main goal of this paper is to paint a comprehensive picture of different weighting approaches. It will reveal which weighting approach leads to the best estimation, i.e., retrieving the true population parameters with least bias and highest precision. Furthermore, we will address the question as to which extent and, why different software packages deliver different results. The aim of the study is to provide a clear recommendation for using weights and estimation procedures for multilevel analyses in LSAs.

This paper is organized as follows. First, we will describe the properties of our example LSA study (PISA) with a focus on its sampling design and weights. Then, different hierarchical models will be introduced in order to obtain a variation of models for the

simulation study. Contextualising the estimation process, the pseudo maximum likelihood estimation method is explained and specifics are discussed. Linking now back to LSAs, different methods for scaling the weights in the hierarchical context are described. Next, the simulation study will be introduced. We explain features of the simulated PISA population, detail sampling-related features, weights and non-response adjustment as well as the analysis procedures. We then present and discuss the results of the simulation study and determine the preferred weighting scheme. This scheme is thereafter applied to the PISA 2015 data (Reiss et al., 2018) with selected hierarchical models. Finally, the results are summarized and possibilities for future research will be discussed.

Methods

PISA sampling design and weights

In all countries participating in PISA, 15-year-old students constitute the target population. In order to collect representative data from this target population in an efficient way, a two-stage sampling design is applied; selecting schools first and students within those schools in a second stage. In preparation of the school sampling, all schools providing education to 15-year-old students are listed using national registers. To make sampling more efficient [i.e., obtain small standard errors (SE)], the whole list of schools is divided into sub-groups, a process called stratification. PISA uses implicit and explicit stratification. Implicit stratification refers to sorting sampling units before sample selection, which is an efficient method to achieve an approximately proportional sample allocation to all strata. Explicit stratification refers to dividing the sampling frame into different groups (in this case, of schools); from each explicit stratum, an independent sample is selected. This stratification method allows disproportional sample allocation (OECD, 2017). For example in Germany, the 16 federal states (explicit stratification) and the different school types (implicit stratification) were used as stratification variables. Within each explicit stratum, schools are now selected using the PPS mechanism, meaning larger schools have a greater probability to be sampled. This selection method leads to significantly varying weights at this first sampling stage. Within every sampled school, 15-year-old students are now randomly sampled as a second selection stage. The within-school sample size, i.e., the number of students to be selected, is settled when defining the target population. In Germany, this target cluster size is, on average over all PISA cycles, approximately 25 students. Mostly, selection probabilities within schools are very similar for all students. To avoid the expected bias due to varying selection probabilities, sampling weights are provided. Those weights are computed as the inverse of the selection probabilities of each selection stage, adjusted for non-response:

$$w_{ij} = w_i * f_{1i} * w'_{ij} * f_{2ij}$$

with w_{ij} as the final student weight for student j in school i , w_i as the base school weight for school i , f_{1i} as the school non-response adjustment, w'_{ij} as the base student weight for student j in school i , f_{2ij} as the student non-response adjustment.

As the school participation is mandatory and therefore the participation rate was over 95% in all previous PISA cycles, adjusting of school non-response has always been minimal in Germany and will be neglected in this paper and the following simulation study.

In PISA, there are three more adjustment factors. Two further correction factors compensate for changes in school size between sampling and data collection. Another correction factor is applied in countries where only 15-year-old students in the class with the highest expected number of 15-year-olds are assessed (OECD, 2017). In the event of non-response at student level, other students who are as similar as possible to the ones who do not participate are given a higher weighting. This avoids under-representation of those students. In detail, non-response adjustment cells are built within each stratum, school, grade and gender (OECD, 2017). This non-response structure is also used in the simulation.

Hierarchical models

In order to be able to represent the variety of hierarchical models, three standard hierarchical models are presented here. Demonstrative and use-oriented examples of all models can be found in Meinck and Vandenplas (2012). For all models, the following notation applies as presented in Table 1.

Model 1—Null model (random intercept)

$$y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$$

Model 2—One explanatory variable at level one with fixed slope (random intercept)

$$y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$$

Model 3—One explanatory variable at level one and level two with fixed slopes (random intercept)

$$y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$$

with $\tau_i \sim N(0, \sigma_\tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$.

Model 1 is technically defined having a school random effect and a residual but no explanatory variable at either level. β_0 is declared as the mean of the achievement. τ_i and ε_{ij} specifies the variance ratio between and within the different levels. Having, for example, an intraclass correlation (ICC) of 0.1 and the students' achievement is given by $\sim N(500, 100)$ the variance is distributed by being 1,000 within the levels and 9,000 between the levels, or in other words only 10% of the variance in achievement is due to school effects. Therefore, this model should be preferred if a researcher is interested in

Table 1 Variable definitions for hierarchical models used in this paper

| | |
|--------------------|--|
| y_{ij} | Student achievement, i.e., PISA competence (Math, Reading or Science) |
| x_{ij} | Student socio-economic status, i.e., the PISA Economic, Social and Cultural Index (ESCS) |
| x_i | The school's socio-economic Index |
| β_0 | Grand (i.e., overall) mean, intercept of the model |
| β_1 | Fixed effect on student level |
| β_2 | Fixed effect on school level |
| ε_{ij} | Residual |
| τ_i | School random effect |

how much of the variance of the dependent variable is determined within and between the levels. As in Model 1, the intercept τ_i in Model 2 is random. The explanatory variable demonstrates a fixed effect to the dependent variable. Researchers should focus on this model if the relation from the independent to the dependent variable at level one after accounting for variation from level two is of interest. Model 3 extends Model 2 by the term $\beta_2 * x_i$ stating the fixed effect of the explanatory variable also at level two.

Pseudo maximum likelihood estimation

In order to enable statistical inference using hierarchical models (i.e., inferring from a sample on an infinite population), two different approaches have been developed, namely design-based and model-based techniques. Design-based methods have their focus on the sample design model with known parameters, assuming, that this model is a true reflection of its population. On the other hand, model-based methods are defining a superpopulation model with unknown parameters having variability from the model error term including that the sample design model is not the superpopulation model (Binder & Roberts, 2010; Snijders & Bosker, 2012).

Asparouhov (2006) and Pfeffermann et al. (1998) defined a hybrid approach combining design-based and model-based inference estimation techniques. The basis is the model-based approach with unknown parameters from the superpopulation model. The focus in this model is not on true parameter estimates, but on estimators, which are design consistent for the infinite population. In conclusion, even if the model assumptions might be wrong, the design consistent estimators are robust. Relating to this hybrid model, the authors note that it is important to include complex sampling designs, like those applied in PISA, in the model. This is done by introducing sampling weights in hierarchical models (Asparouhov, 2006; Graubard & Korn, 1996; Pfeffermann, 1993, 1996). This so called pseudo maximum likelihood (PML) estimation technique was developed by Skinner (1989), following the idea of Binder (1983). Starting with the idea of a model-based approach for reaching statistical inference the census likelihood is defined as

$$L(Y|\theta) = \prod_{j=1}^N f(Y_j|\theta),$$

with $f(Y_j|\theta)$ as the density of Y_j in the population, θ as the unknown population parameter and N the number of students in the population.

To achieve a sum instead of the product for easier mathematical handling, the census log-likelihood follows with

$$l(Y|\theta) = \sum_{j=1}^N \log f(Y_j|\theta).$$

The maximum likelihood (ML) estimate is then obtained by

$$\frac{\partial l(Y|\theta)}{\partial \theta} = 0.$$

Following the hybrid approach stating that the design consistent estimator of the model-based technique is a robust estimator for the infinite population parameters, the principle of the Horvitz–Thompson (HT) estimator is applied (Horvitz & Thompson, 1952; Petkova, 2016). The HT estimator uses the inverse of the selection probabilities as weights

$$\hat{Y}_{HT} = \frac{1}{N} \sum_{j=1}^n w_j y_j = \frac{1}{N} \sum_{j=1}^n \frac{1}{\pi_j} y_j,$$

with π_j as the selection probability, $w_j = \frac{1}{\pi_j}$ as the inverse of the selection probability, y_j as the single characteristics in the sample, N as the population size and n as the sample size.

Transferring this principle to a hierarchical (two level) structure follows the selection probabilities for the schools and students within schools as π_j and π_{ij} , respectively. The weights for the m schools are $w_j = \frac{1}{\pi_j}$ and for the n students $w_{ij} = \frac{1}{\pi_{ij}}$.

Pfeffermann et al. (1998) argued that because of the clustered data structure, observations are not assumed to be independent anymore and the log-likelihood will become a sum across level one and level two elements instead of a simple sum of the element's contributions (Grilli & Pratesi, 2005; Petkova, 2016). Using the idea of the HT estimator with introducing weights into the log-likelihood replaces each sum over the level two population units i by a sample sum weighted by $w_i = \frac{1}{\pi_i}$ and each sum over the level one units j by a sample sum weighted by $w_{ij} = \frac{1}{\pi_{ij}}$ (Grilli & Pratesi, 2005).

The pseudo maximum likelihood estimator $\hat{\theta}_{PML}$ is therefore design consistent for the finite population maximum likelihood estimator $\hat{\theta}$, which, in turn, is model-consistent for the superpopulation estimator of θ . Therefore $\hat{\theta}_{PML}$ is a consistent estimator of θ with respect to the mixed design-model (hybrid) distribution (Pfeffermann et al., 1998).

As no straightforward method of maximising this weighted likelihood function is possible due to the existence of several integrals, numerical approximation techniques can be applied. These optimization techniques will be described in the following passages.

Optimization methods

Historically, the origins of estimating parameters from the weighted likelihood function were located at the so called iterative generalized least squares (IGLS; Goldstein, 1986). This method is based on the normal distribution assumption, implemented and used by Pfeffermann et al. (Pfeffermann & Sverchkov, 2010).

Rabe-Hesketh and Skrondal (2006) choose to solve the weighted likelihood function in the PML equation by using an expectation–maximisation (EM) algorithm (Dempster et al., 1977). The basic idea behind the algorithm is divided into two steps. First, an approximation to the function of interest, i.e., the ML function, with initial, logical parameter values is constructed. This step is called *expectation*. Second, the parameter value, which maximizes this approximation function, is adjusted. This step is named

maximisation. This value is then inserted in the expectation step. The whole procedure is iterated until the parameter values stabilize with a given threshold. Unfortunately, this method suffers from slow convergence rates.

Acceleration methods

Alternative methods to accelerate the EM algorithm are Fisher-Scoring or Quasi-Newton acceleration method. The idea of these methods is not to actually calculate the maximization step of the EM algorithm, but to approximate this calculation. To do that, it takes the so-called score functions, i.e., first and second order derivatives of the approximated ML function, into account (Jamshidian & Jennrich, 1997; Lange, 1995; Longford, 1987). Jamshidian and Jennrich (1997) stated, that these methods accelerated the EM algorithm in some cases by factor 50 and above.

Integration method

In all EM techniques the expectation step is approximated by adaptive quadrature (Bock & Aitkin, 1981). It is a numerical integration method for approximating formulas with integrals. The key is approximating the whole integral by small areas defined by so-called nodes. The principle can be written as

$$\int_a^b f(x) = \sum_{i=1}^n h_i f(x_i),$$

with quadrature nodes $x_i \in [a, b]$, $f(x)$ as any function of interest and quadrature weights h_i which should not be confounded with any weights mentioned in this article. Having a large number of nodes follows a good approximation. Adaptive quadrature places the locations where the integrand is concentrated assuming that the “posteriori” density of a Bayesian perspective is approximately normal distributed (Rabe-Hesketh et al., 2002, 2005).

The SAS[®] software program with its procedure PROC GLIMMIX and its setting adaptive quadrature (SAS Institute Inc., 2018) is based on the EM algorithm estimation *Quasi-Newton* in its default setting, while the Mplus software program (Muthén & Muthén, 2017) declares to use *Fisher-Scoring* in its default setting as accelerated EM method, or also *Quasi-Newton*. The default settings were specified that way to provide also less technical users with a wide range of sophisticated methods.

Sandwich type variance estimation

Besides the estimate itself, its variance (i.e., the squared standard error), is of further interest. The covariance matrix of an estimator is obtained after the model has been estimated. Again, the sampling design needs to be taken into account. If the covariance structure is assumed to be too simple, which is the case for independent random samples, then the model based estimated standard errors for the fixed effects are invalid (usually too small). One way to deal with this is to use sandwich standard errors, which are a function of the modelled standard errors and observed residuals. If the sandwich

standard errors are close to the model-based ones, then one can be confident that the model is well specified. If the model is not correctly specified, then the two types of standard errors will differ, and the sandwich standard errors are preferred. From a technical point of view this variance has been developed by Binder (1983), which is further discussed by Skinner (1989) and is based on Taylor expansion. A general variance estimator is determined by

$$\text{cov}(\hat{\theta}) = K^{-1}JK^{-1}.$$

Here, K is the negative second derivative of the logarithmic pseudo likelihood evaluated at $\hat{\theta}$. In other words, K can be estimated by its empirical mean. The term J designates the estimated variance–covariance matrix of the weighted score functions. It allows taking the sampling weights as well as particular characteristics of the sampling design into account. The crucial point here is the assumption that the residuals of the model are having mean zero (see also “[Hierarchical models](#)” section). Furthermore, the variance is declared as the average squared deviation around the mean. Thus, the estimated residual variance can be written as a sum over schools over students of those squared errors.

This sandwich estimator is implemented by default in most software programs for MLM, including Mplus with its default setting (Muthén & Muthén, 2017) and SAS with its procedure PROC GLIMMIX and its setting for adaptive quadrature (SAS Institute Inc., 2018). Furthermore, there are approaches that specialize in bootstrapping methods. Those methods are used by default in single level LSA analyses (Rust & Rao, 1996).

Scaling methods for level one weights

For most publicly available LSA data sets like PISA, weights for the school level w_i and weights for the student level w_{ij} (“final student weights” combining school and student weights) are provided in order to correctly use weights at each population of interest. Those weights should only be used when analysing data of one population, i.e., either students or schools. Considering more than one level at a time, these weights have to be used or adapted differently in order to account for the hierarchical structure. In other words, including the final student weight w_{ij} would be inappropriate for conducting multilevel analysis (Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006). Pfeffermann et al. (1998) and Rabe-Hesketh and Skrondal (2006) argued further that including unscaled weights in the analysis might lead to bias in the variance estimates. Scaling of level two weights is not considered since it has no effect on the estimates (Bertolet, 2008; Grilli & Pratesi, 2005).

The scaling of level one weights is another approach to take into account the inclusion of weights in hierarchical analyses. Four widely addressed scaling methods are used in the research community, but there is still no clear recommendation which method should be preferred. Furthermore, only the following two methods are (Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006) cited in the literature.

The conditional student weight w_{ij} can be written as

$$w_{ij} = w_{ij}^* \lambda,$$

where λ is a synonym for the scaling factor and w_{ij} defines the weight of student j and school i .

In scaling method 1 the scaled weights add up to the cluster size, i.e., the number of sampled students in a school with $\sum_{j=1}^{n_i} w_{ij}^* = n_i$, so the scale factor can be written as

$$\lambda = \frac{n_i}{\sum_{j=1}^{n_i} w_{ij}^2}.$$

The conditional student weight is then given by

$$w_{ij}^* = w_{ij} \frac{n_i}{\sum_{j=1}^{n_i} w_{ij}},$$

where n_i equals the number of sample units in cluster i . In the simulation study, this method is declared as *Scaled Weights: Cluster*.

In scaling method 2 the sum of the conditional student weights add up to the effective sample size within the cluster, i.e., the number of assessed students in a school with $\sum_{j=1}^{n_i} w_{ij}^* = n_i^*$, so this scale factor can be written as

$$\lambda = \frac{\sum_{j=1}^{n_i} w_{ij}}{\sum_{j=1}^{n_i} w_{ij}^2},$$

and its corresponding conditional student weight as

$$w_{ij}^* = w_{ij} \frac{n_i^*}{\sum_{j=1}^{n_i} w_{ij}}.$$

n_i^* is thereby defined as

$$n_i^* = \frac{\left(\sum_{j=1}^{n_i} w_{ij}\right)^2}{\left(\sum_{j=1}^{n_i} w_{ij}^2\right)}.$$

In the simulation study, this method is declared as *Scaled Weights: ECluster*.

Two further approaches in scaling level one weights are only mentioned in the technical appendixes, but are as often used in analyses as the other approaches. One approach scales the final student weight in order to sum up to the full sample size, given by n . The scale factor can be written as

$$\lambda = \frac{n}{\sum_{j=1}^n w_{ij}^2},$$

and the final scaled student weight as

$$w_{ij}^* = w_{ij} \frac{n}{\sum_{j=1}^n w_{ij}}.$$

This approach is declared as *House Weights* in the simulation study.

The last scaling technique of level one weights described here adds another component to the school weights within the approach *Scaled Weights: Cluster*. Here, the within-school weights add up to the school sample size. Additionally, the school weights are transformed as to reflect the sum of the final student weight within one school given n_i as the number of students within one school. This technique is declared as *Clustersum* in the following simulation study. The transformed school weight can therefore be written as

$$w_i^* = \sum_{j=1}^{n_i} w_{ij}.$$

The most prominent sources presenting this approach are discussed in the below section, “[Analysis procedures](#)”, under Simulation Study. This section also describes the analysis plan.

Research questions

The following research questions will be examined:

- 1) Which weighting scheme performs best in providing population estimates in selected hierarchical models, i.e., with least bias?
- 2) Does scaling of level one weights enhance preciseness and unbiasedness of estimation, and if so, which cited technique should be preferred?
- 3) Which estimation procedure serves for the least biased estimates in selected hierarchical models?

All three research questions are discussed in an independent way, but also considered in combination, because all considered methods are simultaneously at work when conducting analysis with real sample data. The aim of the study is to make a firm proposal for the common estimation of hierarchical models using provided sampling weights.

Simulation study

With the help of a simulation study, the performance of different weighting scenarios within hierarchical models can be investigated by comparing estimated parameters with the true values of a population (Metropolis & Ulam, 1949).

The simulated population mimics the German PISA population. From this “population”, 1000 sample replications are selected according to the population characteristics defined in the next section, using the approach of a Monte Carlo simulation. One thousand replications were considered to be sufficient for achieving stable point estimators (Meinck & Vandenplas, 2012). For each dataset, simulated weights are calculated when drawing the sample.

The software program R Studio Version 1.1.456 (RStudio Team, 2018) and its corresponding program R 3.5.1 (R Core Team, 2018) was used for simulating the sample replicates. The analyses was performed with two software programs for hierarchical modelling of large-scale assessment data Mplus (Muthén & Muthén, 2017) and

SAS with its procedure PROC GLIMMIX (SAS Institute Inc., 2018). Both software packages are widely used in the researcher community, especially among educational researchers, and of special interest for the authors. Three representative hierarchical models were analysed.

Simulation PISA population

The simulation of the population of 15-year-old students is based on two data sources. The first source was the sampling frame for PISA 2015 in Germany. In this frame, all schools accommodating 15-year-old students in the school year 2012/2013 are listed, together with their allocation to federal state and school type, and the expected number of 15-year-old students. Information originates from federal and governmental offices. Further, relevant population features were estimated based on the German PISA 2015 sample and added to each school on the above-mentioned sampling frame.

In order to investigate the differential effects of varying parameters, three different simulation scenarios for generating the student achievement data (i.e., the PISA competence for a given domain) and socio-economic background were implemented.

For the first scenario, the population parameters are chosen in a way to correspond to the true German PISA target population in 2015. To achieve this, real outcomes of the PISA 2015 cycle were used. That is, the performance in science (first PV) and the PISA Economic, Social and Cultural Index (ESCS) for the socio-economic index split for each different school type served as scenario templates (Simulation Scenario 1).

Secondly, a scenario with nearly no variance between the schools of a given school type is simulated (Simulation Scenario 2). The ICC of 0.05 is very small in this scenario, and MLM may not be that advantageous to single-level analysis under such circumstances. We still decided to implement such scenario for two reasons. One was to get a good contrast for the scenarios with higher ICC. Second, some authors (e.g. Snijders & Bosker, 2012) recommend MLM whenever there is a hierarchical structure in the underlying population. Also Lai and Kwok (2015) recommend hierarchical modelling in such scenarios because there is in fact still a design effect (Kish, 1965) to account for.

The third scenario is based on a high variance between the schools of a given school type (Simulation Scenario 3). All simulation scenarios comprise a two-level structure with schools at level one and students at level two.

For each of the three scenarios, the different compositions of the performance of the schools (i.e., the school achievement) and their socio-economic index were simulated. Following this, the performance and socio-economic status of each student was simulated around those school values, with a given variance and covariance according to the appropriate simulation scenario. Overall, 16,330 schools and 841,095 students are simulated for each single simulation scenario.

Table 2 shows the different simulation scenarios and their corresponding characteristics. As population parameters for one scenario can vary between the three chosen hierarchical models, those values are indicated with “/” for each model within the appropriate scenario. For variable definitions please refer to Tables 1.

Table 2 Population specifications

| Population Parameters | Scenario 1—PISA | Scenario 2—low | Scenario 3—high |
|-----------------------|--------------------|-------------------|-----------------------|
| y_{ij} | $\sim N(505, 101)$ | $\sim N(500, 97)$ | $\sim N(468, 148)$ |
| x_{ij} | $\sim N(0, 1)$ | $\sim N(0, 0.89)$ | $\sim N(0.16, 1.20)$ |
| x_i | $\sim N(0, 0.59)$ | $\sim N(0, 0.36)$ | $\sim N(-0.10, 0.89)$ |
| β_0 | 476/ 479/ 494 | 500/500/500 | 421/429/449 |
| β_1 | -/29/28 | -/27/26 | -/35/35 |
| β_2 | -/-/40 | -/-/7 | -/-/65 |
| ε_{ij} | 5005/4022/4027 | 8994/8421/8419 | 5012/4420/4420 |
| τ_i | 5053/4240/ 2417 | 530/299/266 | 16,100/12,541/8191 |
| ICC | 0.52 | 0.05 | 0.79 |

Samples, weights and non-response

The federal states and the school types served as explicit and implicit stratification variables in Germany (OECD, 2017). There are 16 federal states. The different school types comprise lower secondary, upper secondary and vocational schools with basic or advanced general educational tracks. Explicit stratification implies that schools are sampled independently for each stratum. Mirroring the sampling procedure from 2015, we divided the sampling frame by federal states, and then sorted schools within states by type and their expected numbers of 15-year-old students. In the next step, 1000 samples of 234 schools with a maximum of 25 students per school were drawn by PPS sampling for each simulation scenario. Two hundred and thirty-four schools are chosen to satisfy minimum sample size requirements for explicit strata in PISA 2015. In schools with less than 25 eligible students, all of them were selected.

Sampling weights applied in PISA reflect the PPS sampling technique that leads to approximately self-weighted samples (Särndal et al., 2003). Larger schools have a higher probability to be selected whereas students in these schools have smaller probabilities to be part of the sample. PPS sampling applied in PISA leads to similar final student weights, but to school base weights that follow a Poisson distribution (Särndal et al., 2003). The school base weights as well as the student base weights can be generated directly when drawing the school and the student sample. The full student base weight as a product over the school and the student base weight is then given by

$$w_{ij} = \frac{1}{\pi_{ij}},$$

with π_{ij} is the selection probability for student j in school i .

In order to achieve the final school and student weights, non-response for both levels must be considered. As the assessment is mandatory in Germany, non-response for schools was very low over most cycles, hence we assumed 100% participation at school level for the simulation. The three further adjustment factors mentioned earlier are equal to one in the vast majority of cases over all cycles, therefore they are neglected as well in the simulation study. At the student level, non-response is simulated similar to PISA procedures. Combined non-response is adjusted by grade and gender characteristics (OECD, 2017). A logistic regression model generates student

participating probability weights, which are dependent on the student's gender and grade. As the distribution of girls and boys participating in PISA is nearly 50/50, this proportion is kept for the simulation study. The modal grade in PISA 2015 and therefore used for this simulation was given by nearly 50% in grade 9 and 50% in grade 10. Only a very limited number of PISA students attend grades 7, 8 or 11, so this portion is neglected. The regression model for simulating student non-response is thus given by

$$\log(P(Y_{ij} = 1)) = \beta_0 + \beta_1 * gender_{ij} + \beta_2 * grade_{ij},$$

with $\beta_0 = 0.1$, $\beta_1 = \beta_2 = 0.05$, $Y_{ij} \in [0, 1]$, $gender_{ij} \in [0, 1]$ and $grade_{ij} \in [0, 1]$.

A uniform random sample determines if a student is set to participating or non-responding. This participating probability is then distributed across participating students.

Analysis procedures

Table 3 shows the different weighting scenarios combined with different software programs and estimation methods applied in the simulation study. All simulation scenarios and weighting approaches are applied to each hierarchical model explained in “[Methods](#)” section (Table 3).

Overall, 126 different scenarios have been analysed, each with 1000 replications using the Monte Carlo approach. It was deemed that 1000 repetitions were sufficient to achieve stable and highly precise estimates of model parameters and their SEs (Meinck & Vandenplas, 2012). A nearly exact representation of the target population becomes possible, so that estimates can be reliably compared with the true population values.

Nine different weighting approaches were selected to provide a comprehensive and nearly complete picture of all possible variants. The following table shows all approaches and their application to the different levels of the hierarchies (Table 4).

The weighting scenario *No Weights* at both levels stands for no weighting at either school or student level. The approach *Unscaled Weights* at both levels uses both weights, i.e., the school weight and the final student weight at each level. The scenario *Only Student Weights* and *Only School Weights* each weight at the respective level only. The school weight represents the inverse of the school selection probability, adjusted for school nonresponse. The student weight equals to the final student weight in this scenario.

Scenario *House Weights* reflects the approach of scaling the final student weights to sum up to the sample size. Former PISA analyses and recommendations (OECD, 2009) as well as former MLM analysis based on TIMSS and PIRLS refer to this procedure (Martin & Mullis, 2013).

Using school weights at level two and scaled student weights at level one with different scaling techniques is implemented in the approaches *Cluster* and *Ecluster*, each based on the appropriate scaling explained in the section Scaling Methods for Level One Weights. Multilevel analyses in the PISA 2009 report volume VI (OECD, 2011) use this approach. Since the PISA 2012 cycle, the OECD is following another approach, here named *Clustersum*. In this approach, the within-school weights are also scaled to sum up to the cluster sample size (as in the approach *Cluster*), but school weights are handled to reflect the

Table 3 Simulation scenarios including varying ICCs, three investigated hierarchical models and different weighting approaches combined with different estimation algorithms implemented in the two examined software packages

| ICC | Model | Software package | Weighting scenario |
|----------------|---------|------------------|--------------------------|
| 0.52/0.05/0.79 | Model 1 | MPLUS | No weights |
| | | | Unscaled weights |
| | | | Only student weights |
| | | | Only school weights |
| | | | Scaled weights: cluster |
| | | | Scaled weights: ECluster |
| | | | Withincluster weights |
| | | | House weights |
| | | | Clustersum |
| | | SAS | No weights |
| | | | Unscaled weights |
| | | | Only student weights |
| | | | Only school weights |
| | | | Scaled weights: cluster |
| | | | Scaled weights: ECluster |
| | | | Withincluster weights |
| | | | House weights |
| | | | Clustersum |
| | Model 2 | MPLUS | No weights |
| | | | Unscaled weights |
| | | | Only student weights |
| | | | Only school weights |
| | | | Scaled weights: cluster |
| | | | Scaled weights: ECluster |
| | | | Withincluster Weights |
| | | | House weights |
| | | | Clustersum |
| | | SAS | No weights |
| | | | Unscaled weights |
| | | | Only student weights |
| | | | Only school weights |
| | | | Scaled weights: cluster |
| | | | Scaled weights: ECluster |
| | | | Withincluster weights |

Table 3 (continued)

| ICC | Model | Software package | Weighting scenario |
|-----|---------|------------------|--------------------------|
| | Model 3 | MPLUS | House weights |
| | | | Clustersum |
| | | | No weights |
| | | | Unscaled Weights |
| | | | Only student weights |
| | | | Only school weights |
| | | | Scaled weights: cluster |
| | | | Scaled weights: ECluster |
| | | | Withincluster weights |
| | | | House weights |
| | | SAS | Clustersum |
| | | | No weights |
| | | | Unscaled weights |
| | | | Only student weights |
| | | | Only school weights |
| | | | Scaled weights: cluster |
| | | | Scaled weights: ECluster |
| | | | Withincluster weights |
| | | | House weights |
| | | | Clustersum |

Model 1 is declared as $y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$, Model 2 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$ and Model 3 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$

sum of the final student weights within one school. The authors claim this approach is more student-centred (OECD, 2014, 2016, 2019).

The approach *Withincluster Weights* applies school weights at level two, and at level one the inverse of the selection probability of a student within a school, adjusted for non-response. The school weights are only included at school level and not as an additional factor in the full student weights. This scenario focuses on the respective adjustments that are assigned to the hierarchical levels and refers to Rutkowski et al. (2010). The International Civic and Citizenship Education Study (ICCS) (Schulz et al., 2018) and the International Computer and Information Literacy Study (ICILS) (Gebhardt et al., 2014) implemented this approach.

All analyses were performed using Mplus Version 8.1 (Muthén & Muthén, 2017) and SAS Version 9.4 (SAS Institute Inc., 2018) with its procedure PROC GLIMMIX.

Table 4 Weighting approaches for the simulation study and their application and formulas for the different levels of the hierarchies

| Weighting approaches | School level | Student level |
|--------------------------|---------------------------|--|
| No weights | — | — |
| Unscaled weights | w_i | w_{ij} |
| Only student weights | | w_{ij} |
| Only school weights | w_i | |
| Scaled weights: Cluster | w_i | $w_{ij} \frac{n_i}{\sum_{j=1}^{n_i} w_{ij}}$ |
| Scaled weights: ECluster | w_i | $w_{ij} \frac{n_i^*}{\sum_{j=1}^{n_i} w_{ij}}$ |
| Withincluster weights | w_i^* | w_{ij}^* |
| House weights | | $w_{ij} \frac{n}{\sum_{j=1}^n w_{ij}}$ |
| Clustersum | $\sum_{j=1}^{n_i} w_{ij}$ | $w_{ij} \frac{n_i}{\sum_{j=1}^{n_i} w_{ij}}$ |

Weighting parameters are w_i = final school weights, w_{ij} = final student weights, n_i = number of sampled students in a school, n_i^* = number of assessed students in a school, n = number of assessed students from all schools and w_j = final within school weights

Results and discussion

In the following, figures of boxplots to the estimation parameters from the respective chosen model are displayed. *Boxplots* describe the distribution of an estimated value based on many repetitions (1000 in our study). The median, the 25% and 75% quartiles, minimum and maximum are presented (Chambers, 1983). Differences between the boxplots are interpreted based on several definitions (e.g. Williamson et al., 1989). Firstly, the boxes representing the interquartile ranges are compared. If boxes do not overlap, a difference can be stated. Secondly, medians are considered. If the median line of a box lies outside of another box entirely, then a difference between the two groups is likely. Thirdly, the whiskers must be considered. They mark the maximum and the minimum values of each set. Their distance represents the range between those two extremes. Larger ranges indicate wider distribution, that is, more scattered data. Since differences in the boxplots between the various weighting approaches can usually already be determined based on the median deviations and the interquartile distances, the whiskers are barely discussed below. In addition to the graphical results, empirical 95% coverage rates (CR) for each parameter are given in Tables 5, 6 and 7 for each simulation scenario, respectively. The empirical 95% coverage rate indicates how often the 95% confidence interval of each estimated parameter covers the true population value. A good coverage rate starts at 95%.

Figure 1 shows the three selected hierarchical models based on the simulation of the PISA data (Simulation Scenario 1). Figure 2 refers to the Simulation Scenario 2 with low variances between the schools and Fig. 3 refers to Simulation Scenario 3 with high variances between those schools. The figures present the estimated fixed parameters as well as the estimated variances within and between the schools for each model in the appropriate simulation scenario. The true population values for each estimate are marked as red line in each graph. The closer the boxplot median line to the red line, the better does the respective estimation method retrieve the true population parameter. If the box does not cover the true population value, the estimation is highly biased. The larger the box, the less precise is the estimation method. When comparing results between the software

Table 5 Coverage Rates of PISA simulated data

| Software | Weighting approach | CR $\hat{\beta}_0$ | CR $\hat{\beta}_1$ | CR $\hat{\beta}_2$ | CR $\hat{\sigma}_\varepsilon^2$ | CR $\hat{\sigma}_\tau^2$ |
|---|--------------------------|--------------------|--------------------|--------------------|---------------------------------|--------------------------|
| A: Coverage rates—PISA simulated data—Model 1 | | | | | | |
| SAS | No weights | 0.00 | | | 0.94 | 0.98 |
| | Unscaled weights | 1.00 | | | 0.51 | 0.99 |
| | Only school weights | 1.00 | | | 0.94 | 0.95 |
| | Only student weights | 0.00 | | | 0.32 | 0.96 |
| | Withincluster weights | 1.00 | | | 0.62 | 1.00 |
| | Scaled weights: cluster | 1.00 | | | 0.95 | 0.95 |
| | Scaled Weights: ECluster | 1.00 | | | 0.95 | 0.95 |
| | Clustersum | 0.00 | | | 0.95 | 0.99 |
| | House weights | 0.00 | | | 0.94 | 0.99 |
| | | | | | | |
| Mplus | No weights | 0.00 | | | 0.92 | 1.00 |
| | Unscaled weights | 1.00 | | | 0.97 | 0.97 |
| | Only school weights | 1.00 | | | 0.97 | 0.97 |
| | Only student weights | 0.00 | | | 0.92 | 1.00 |
| | Withincluster weights | 1.00 | | | 0.97 | 0.97 |
| | Scaled weights: cluster | 1.00 | | | 0.96 | 0.96 |
| | Scaled weights: ECluster | 1.00 | | | 0.96 | 0.96 |
| | Clustersum | 0.00 | | | 0.97 | 1.00 |
| | House weights | 0.00 | | | 0.97 | 1.00 |
| | | | | | | |
| B: Coverage rates—PISA simulated data—Model 2 | | | | | | |
| SAS | No weights | 0.00 | 0.83 | | 0.94 | 0.00 |
| | Unscaled weights | 0.99 | 0.90 | | 0.53 | 0.00 |
| | Only school weights | 0.98 | 0.91 | | 0.95 | 0.84 |
| | Only student weights | 0.00 | 0.85 | | 0.37 | 0.00 |
| | Withincluster weights | 0.98 | 0.94 | | 0.68 | 0.59 |
| | Scaled weights: cluster | 0.99 | 0.90 | | 0.96 | 0.81 |
| | Scaled weights: ECluster | 0.99 | 0.90 | | 0.96 | 0.81 |
| | Clustersum | 0.00 | 0.92 | | 0.96 | 0.39 |
| | House weights | 0.00 | 0.85 | | 0.94 | 0.00 |
| | | | | | | |
| Mplus | No weights | 0.00 | 0.91 | | 0.95 | 0.61 |
| | Unscaled weights | 0.98 | 0.92 | | 0.94 | 0.96 |
| | Only school weights | 0.98 | 0.92 | | 0.95 | 0.96 |
| | Only student weights | 0.00 | 0.91 | | 0.95 | 0.61 |
| | Withincluster weights | 0.98 | 0.92 | | 0.94 | 0.96 |
| | Scaled weights: cluster | 0.99 | 0.91 | | 0.95 | 0.97 |
| | Scaled weights: ECluster | 0.99 | 0.91 | | 0.95 | 0.97 |
| | Clustersum | 0.00 | 0.92 | | 0.96 | 0.39 |
| | House weights | 0.00 | 0.92 | | 0.95 | 0.65 |
| | | | | | | |
| C: Coverage rates—PISA simulated data—Model 3 | | | | | | |
| SAS | No weights | 0.01 | 0.93 | 0.96 | 0.94 | 0.44 |
| | Unscaled weights | 0.96 | 0.94 | 0.93 | 0.53 | 0.45 |
| | Only school weights | 0.95 | 0.94 | 0.93 | 0.94 | 0.87 |
| | Only student weights | 0.09 | 0.94 | 0.93 | 0.36 | 0.23 |
| | Withincluster weights | 0.94 | 0.94 | 0.92 | 0.68 | 0.82 |
| | Scaled weights: Cluster | 0.96 | 0.93 | 0.93 | 0.95 | 0.86 |
| | Scaled weights: ECluster | 0.96 | 0.93 | 0.93 | 0.95 | 0.86 |
| | Clustersum | 0.01 | 0.92 | 0.96 | 0.96 | 0.38 |
| | House weights | 0.03 | 0.92 | 0.95 | 0.95 | 0.47 |
| | | | | | | |

Table 5 (continued)

| Software | Weighting approach | CR $\hat{\beta}_0$ | CR $\hat{\beta}_1$ | CR $\hat{\beta}_2$ | CR $\hat{\sigma}_e^2$ | CR $\hat{\sigma}_\tau^2$ |
|----------|--------------------------|--------------------|--------------------|--------------------|-----------------------|--------------------------|
| Mplus | No weights | 0.01 | 0.93 | 0.96 | 0.94 | 0.52 |
| | Unscaled weights | 0.95 | 0.94 | 0.93 | 0.95 | 0.91 |
| | Only school weights | 0.95 | 0.94 | 0.92 | 0.95 | 0.91 |
| | Only student weights | 0.01 | 0.93 | 0.96 | 0.94 | 0.51 |
| | Withincluster weights | 0.95 | 0.94 | 0.93 | 0.95 | 0.91 |
| | Scaled weights: cluster | 0.96 | 0.93 | 0.93 | 0.96 | 0.9 |
| | Scaled weights: ECluster | 0.96 | 0.93 | 0.93 | 0.96 | 0.9 |
| | Clustersum | 0.01 | 0.92 | 0.96 | 0.96 | 0.38 |
| | House Weights | 0.03 | 0.92 | 0.95 | 0.95 | 0.51 |

The CR represents the compliance rate of the estimators within its 95% confidence interval of three hierarchical models. Model 1 is declared as $y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$, Model 2 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$ and Model 3 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$. PISA simulated data serves as scenario template. Simulation variation is displayed with the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages

packages SAS and Mplus, we consistently refer to the software settings specified earlier (SAS: procedure PROC GLIMMIX and its setting adaptive quadrature; Mplus: default settings for two-level modelling).

Outcomes for simulation scenario 1 (data mirroring the German PISA population)

Model 1

It can be seen in Fig. 1, Graph A, Graph D and Graph H, that in all three models the weighting approaches *No Weights*, *Only Student Weights*, *Clustersum* and *House Weights* overestimate drastically the intercept $\hat{\beta}_0$ as the respective boxes do not cover the true population value. Furthermore, medians do not even come close to the true value. This can also be confirmed by looking at the coverage rates of 0% in Table 5 A $\hat{\beta}_0$. This result reflects the German PISA sample structure, where small schools have low selection probabilities and at the same time systematically lower average achievement than large schools (with high selection probabilities), as many of them accommodate students with special educational needs or vocational students. When neglecting school weights, these parts of the target population are underrepresented, which explains the overestimated average achievement. This result provides solid evidence to generally recommend the use of school weights in hierarchical models.

Looking at the next model parameter, we can see that Fig. 1, Graph B, the weighting approaches *Unscaled Weights*, *Only Student Weights* and *Withincluster Weights* underestimate the *Variance Within* $\hat{\sigma}_e^2$ the schools, if using the software program SAS for estimation. This occurs also with all three hierarchical models (Fig. 1, Graph F and Graph K). The weighting approaches *No Weights* (for both software programs), *Only Student Weights* (for both software programs), *Unscaled Weights* (for the software program SAS), *Clustersum* (for both software programs) and *House Weights* (for both software programs) underestimate the *Variance Between* $\hat{\sigma}_\tau^2$ of the

Table 6 Coverage rates of low variances simulated data

| Software | Weighting approach | CR $\hat{\beta}_0$ | CR $\hat{\beta}_1$ | CR $\hat{\beta}_2$ | CR $\hat{\sigma}_\varepsilon^2$ | CR $\hat{\sigma}_\tau^2$ |
|--|--------------------------|--------------------|--------------------|--------------------|---------------------------------|--------------------------|
| A: Coverage rates—low variances simulated data—Model 1 | | | | | | |
| SAS | No weights | 0.94 | | | 0.94 | 0.87 |
| | Unscaled weights | 0.93 | | | 0.62 | 0.74 |
| | Only school weights | 0.94 | | | 0.94 | 0.87 |
| | Only student weights | 0.91 | | | 0.43 | 0.70 |
| | Withincluster weights | 0.94 | | | 0.77 | 0.94 |
| | Scaled weights: cluster | 0.94 | | | 0.94 | 0.87 |
| | Scaled weights: ECluster | 0.94 | | | 0.94 | 0.87 |
| | Clustersum | 0.94 | | | 0.95 | 0.87 |
| | House weights | 0.94 | | | 0.95 | 0.90 |
| | | | | | | |
| Mplus | No weights | 0.94 | | | 0.94 | 0.92 |
| | Unscaled weights | 0.95 | | | 0.94 | 0.90 |
| | Only school weights | 0.94 | | | 0.94 | 0.90 |
| | Only student weights | 0.94 | | | 0.94 | 0.91 |
| | Withincluster weights | 0.94 | | | 0.94 | 0.91 |
| | Scaled weights: cluster | 0.95 | | | 0.94 | 0.90 |
| | Scaled weights: ECluster | 0.95 | | | 0.94 | 0.90 |
| | Clustersum | 0.94 | | | 0.94 | 0.92 |
| | House weights | 0.94 | | | 0.94 | 0.91 |
| | | | | | | |
| B: Coverage rates—low variances simulated data—Model 2 | | | | | | |
| SAS | No weights | 0.89 | 0.90 | | 0.94 | 0.91 |
| | Unscaled weights | 0.91 | 0.92 | | 0.62 | 0.06 |
| | Only school weights | 0.91 | 0.91 | | 0.96 | 0.91 |
| | Only student weights | 0.87 | 0.91 | | 0.41 | 0.72 |
| | Withincluster weights | 0.91 | 0.93 | | 0.78 | 0.52 |
| | Scaled weights: cluster | 0.91 | 0.92 | | 0.96 | 0.91 |
| | Scaled weights: ECluster | 0.91 | 0.92 | | 0.96 | 0.91 |
| | Clustersum | 0.89 | 0.91 | | 0.95 | 0.92 |
| | House weights | 0.90 | 0.91 | | 0.95 | 0.93 |
| | | | | | | |
| Mplus | No weights | 0.89 | 0.90 | | 0.95 | 0.92 |
| | Unscaled weights | 0.91 | 0.92 | | 0.95 | 0.92 |
| | Only school weights | 0.91 | 0.92 | | 0.95 | 0.92 |
| | Only student weights | 0.89 | 0.90 | | 0.95 | 0.93 |
| | Withincluster weights | 0.91 | 0.92 | | 0.95 | 0.92 |
| | Scaled weights: cluster | 0.91 | 0.92 | | 0.95 | 0.92 |
| | Scaled weights: ECluster | 0.91 | 0.92 | | 0.95 | 0.92 |
| | Clustersum | 0.89 | 0.91 | | 0.95 | 0.92 |
| | House weights | 0.89 | 0.90 | | 0.95 | 0.93 |
| | | | | | | |
| C: Coverage rates—low variances simulated data—Model 3 | | | | | | |
| SAS | No weights | 0.87 | 0.93 | 0.95 | 0.95 | 0.91 |
| | Unscaled weights | 0.91 | 0.93 | 0.93 | 0.62 | 0.03 |
| | Only school weights | 0.90 | 0.94 | 0.95 | 0.96 | 0.90 |
| | Only student weights | 0.86 | 0.93 | 0.94 | 0.40 | 0.52 |
| | Withincluster weights | 0.90 | 0.93 | 0.95 | 0.74 | 0.43 |
| | Scaled weights: cluster | 0.90 | 0.94 | 0.95 | 0.96 | 0.91 |
| | Scaled weights: ECluster | 0.90 | 0.94 | 0.95 | 0.96 | 0.90 |
| | Clustersum | 0.87 | 0.93 | 0.94 | 0.95 | 0.91 |
| | House weights | 0.88 | 0.93 | 0.95 | 0.95 | 0.92 |
| | | | | | | |

Table 6 (continued)

| Software | Weighting approach | CR $\hat{\beta}_0$ | CR $\hat{\beta}_1$ | CR $\hat{\beta}_2$ | CR $\hat{\sigma}_\varepsilon^2$ | CR $\hat{\sigma}_\tau^2$ |
|----------|--------------------------|--------------------|--------------------|--------------------|---------------------------------|--------------------------|
| Mplus | No weights | 0.87 | 0.93 | 0.95 | 0.95 | 0.93 |
| | Unscaled weights | 0.90 | 0.94 | 0.95 | 0.95 | 0.92 |
| | Only school weights | 0.90 | 0.94 | 0.95 | 0.95 | 0.91 |
| | Only student weights | 0.87 | 0.93 | 0.95 | 0.95 | 0.93 |
| | Withincluster weights | 0.90 | 0.94 | 0.95 | 0.95 | 0.92 |
| | Scaled weights: Cluster | 0.90 | 0.94 | 0.95 | 0.95 | 0.92 |
| | Scaled weights: Ecluster | 0.90 | 0.94 | 0.95 | 0.95 | 0.92 |
| | Clustersum | 0.87 | 0.93 | 0.94 | 0.95 | 0.93 |
| | House weights | 0.87 | 0.93 | 0.95 | 0.95 | 0.93 |

The CR represents the compliance rate of the estimators within its 95% confidence interval of three hierarchical models. Model 1 is declared as $y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$, Model 2 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$ and Model 3 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$. Low variances between schools simulated data serves as scenario template. Simulation variation is displayed with the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages

schools (Fig. 1, Graph G and Graph L). Interestingly, for Model 1 (Fig. 1, Graph C), the *Variance Between* $\hat{\sigma}_\tau^2$ seems to be overestimated throughout nearly all weighting scenarios when using the software package Mplus as none of the boxplots cover the true value. These facts are also reflected in the coverage rates in Table 5 A $\hat{\sigma}_\tau^2$. Both software programs use the sandwich type estimator for calculating standard errors in the hierarchical models, which is based on the sampling weights, particular characteristics of the sampling design as well as the maximum likelihood function of the appropriate model. As both software packages SAS and Mplus are not as transparent as freely available software packages like R (R Core Team, 2018), we can only guess what distinguishes the two software programs. For example, different accelerating methods for optimization could cause the differences.

Models 2 and 3

By adding the socio-economic background regressor at the student level in Model 2 (Fig. 1, Graph E), it becomes evident that the weighting approaches *Unscaled Weights*, *Only Student Weights* and *Withincluster Weights* also slightly underestimate this estimator $\hat{\beta}_1$ by the SAS software program with its procedure GLIMMIX although inter-quartile spaces include the true value and overlap with one another. However, this effect is offset by the addition of the average SES $\hat{\beta}_2$ at school level in Model 3 (Fig. 1, Graph I and Graph J). From Model 1 to Model 2 (Fig. 1, Graph B and Graph F), the *Variance Within* the schools $\hat{\sigma}_\varepsilon^2$ decreases. This is caused by the increase in explained variance by adding the SES indicator. The same applies for the *Variance Between* the schools $\hat{\sigma}_\tau^2$ as it decreases from Model 2 to Model 3 (Fig. 1, Graph G and Graph L).

Since proposals for weighting approaches working independently of the selected software programs would be desirable, only three weighting approaches provide sufficiently unbiased estimates in this simulation scenario: *Only School Weights*, *Scaled Weights: Cluster* and *Scaled Weights: Ecluster*. All three of these approaches perform nearly the same, as can be seen by having a closer look at their coverage rates

Table 7 Coverage rates of high variances simulated data

| Software | Weighting approach | CR $\hat{\beta}_0$ | CR $\hat{\beta}_1$ | CR $\hat{\beta}_2$ | CR $\hat{\sigma}_\epsilon^2$ | CR $\hat{\sigma}_\tau^2$ |
|---|--------------------------|--------------------|--------------------|--------------------|------------------------------|--------------------------|
| A: Coverage rates—high variances simulated data—Model 1 | | | | | | |
| SAS | No weights | 0.00 | | | 0.93 | 0.97 |
| | Unscaled weights | 0.98 | | | 0.52 | 0.99 |
| | Only school weights | 0.99 | | | 0.94 | 0.99 |
| | Only student weights | 0.00 | | | 0.32 | 0.97 |
| | Withincluster weights | 0.99 | | | 0.66 | 0.99 |
| | Scaled weights: cluster | 0.99 | | | 0.94 | 0.99 |
| | Scaled weights: ECluster | 0.99 | | | 0.94 | 0.99 |
| | Clustersum | 0.00 | | | 0.94 | 0.96 |
| | House weights | 0.00 | | | 0.93 | 0.96 |
| | | | | | | |
| Mplus | No weights | 0.00 | | | 0.94 | 0.90 |
| | Unscaled weights | 0.99 | | | 0.94 | 0.97 |
| | Only school weights | 0.99 | | | 0.94 | 0.97 |
| | Only student weights | 0.00 | | | 0.94 | 0.90 |
| | Withincluster weights | 0.99 | | | 0.94 | 0.97 |
| | Scaled weights: cluster | 0.99 | | | 0.95 | 0.98 |
| | Scaled Weights: ECluster | 0.99 | | | 0.95 | 0.98 |
| | Clustersum | 0.00 | | | 0.95 | 0.78 |
| | House weights | 0.00 | | | 0.94 | 0.82 |
| | | | | | | |
| B: Coverage rates—high variances simulated data—Model 2 | | | | | | |
| SAS | No weights | 0.00 | 0.84 | | 0.94 | 0.02 |
| | Unscaled weights | 0.99 | 0.89 | | 0.54 | 0.06 |
| | Only school weights | 0.99 | 0.86 | | 0.94 | 0.07 |
| | Only student weights | 0.00 | 0.82 | | 0.35 | 0.02 |
| | Withincluster weights | 0.99 | 0.88 | | 0.71 | 0.06 |
| | Scaled weights: cluster | 0.99 | 0.86 | | 0.95 | 0.07 |
| | Scaled weights: ECluster | 0.99 | 0.86 | | 0.95 | 0.07 |
| | Clustersum | 0.00 | 0.89 | | 0.94 | 0.96 |
| | House weights | 0.00 | 0.84 | | 0.95 | 0.02 |
| | | | | | | |
| Mplus | No weights | 0.00 | 0.88 | | 0.94 | 0.97 |
| | Unscaled weights | 0.98 | 0.89 | | 0.93 | 0.94 |
| | Only school weights | 0.98 | 0.90 | | 0.94 | 0.94 |
| | Only student weights | 0.00 | 0.88 | | 0.94 | 0.97 |
| | Withincluster weights | 0.98 | 0.89 | | 0.93 | 0.94 |
| | Scaled weights: cluster | 0.99 | 0.88 | | 0.94 | 0.95 |
| | Scaled weights: ECluster | 0.99 | 0.89 | | 0.94 | 0.94 |
| | Clustersum | 0.00 | 0.89 | | 0.94 | 0.96 |
| | House weights | 0.00 | 0.87 | | 0.95 | 0.96 |
| | | | | | | |
| C: Coverage rates—high variances simulated data—Model 3 | | | | | | |
| SAS | No weights | 0.12 | 0.90 | 0.96 | 0.94 | 0.96 |
| | Unscaled weights | 0.96 | 0.89 | 0.93 | 0.56 | 0.99 |
| | Only school weights | 0.96 | 0.91 | 0.94 | 0.93 | 0.99 |
| | Only student weights | 0.13 | 0.90 | 0.96 | 0.37 | 0.95 |
| | Withincluster weights | 0.96 | 0.90 | 0.94 | 0.71 | 0.99 |
| | Scaled weights: cluster | 0.98 | 0.92 | 0.94 | 0.94 | 0.98 |
| | Scaled weights: ECluster | 0.97 | 0.92 | 0.94 | 0.94 | 0.98 |
| | Clustersum | 0.10 | 0.89 | 0.95 | 0.94 | 0.94 |
| | House weights | 0.13 | 0.88 | 0.95 | 0.94 | 0.93 |
| | | | | | | |

Table 7 (continued)

| Software | Weighting approach | CR $\hat{\beta}_0$ | CR $\hat{\beta}_1$ | CR $\hat{\beta}_2$ | CR $\hat{\sigma}_\varepsilon^2$ | CR $\hat{\sigma}_\tau^2$ |
|----------|--------------------------|--------------------|--------------------|--------------------|---------------------------------|--------------------------|
| Mplus | No weights | 0.11 | 0.90 | 0.96 | 0.94 | 0.96 |
| | Unscaled weights | 0.96 | 0.91 | 0.94 | 0.93 | 0.92 |
| | Only school weights | 0.96 | 0.90 | 0.94 | 0.93 | 0.92 |
| | Only student weights | 0.11 | 0.91 | 0.96 | 0.94 | 0.96 |
| | Withincluster weights | 0.96 | 0.91 | 0.94 | 0.93 | 0.92 |
| | Scaled weights: cluster | 0.97 | 0.91 | 0.94 | 0.94 | 0.92 |
| | Scaled Weights: Ecluster | 0.97 | 0.91 | 0.94 | 0.94 | 0.92 |
| | Clustersum | 0.10 | 0.89 | 0.94 | 0.94 | 0.94 |
| | House weights | 0.14 | 0.89 | 0.95 | 0.94 | 0.94 |

The CR represents the compliance rate of the estimators within its 95% confidence interval of three hierarchical models. Model 1 is declared as $y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$, Model 2 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$ and Model 3 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$. High variances between schools simulated data serves as scenario template. Simulation variation is displayed with the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages

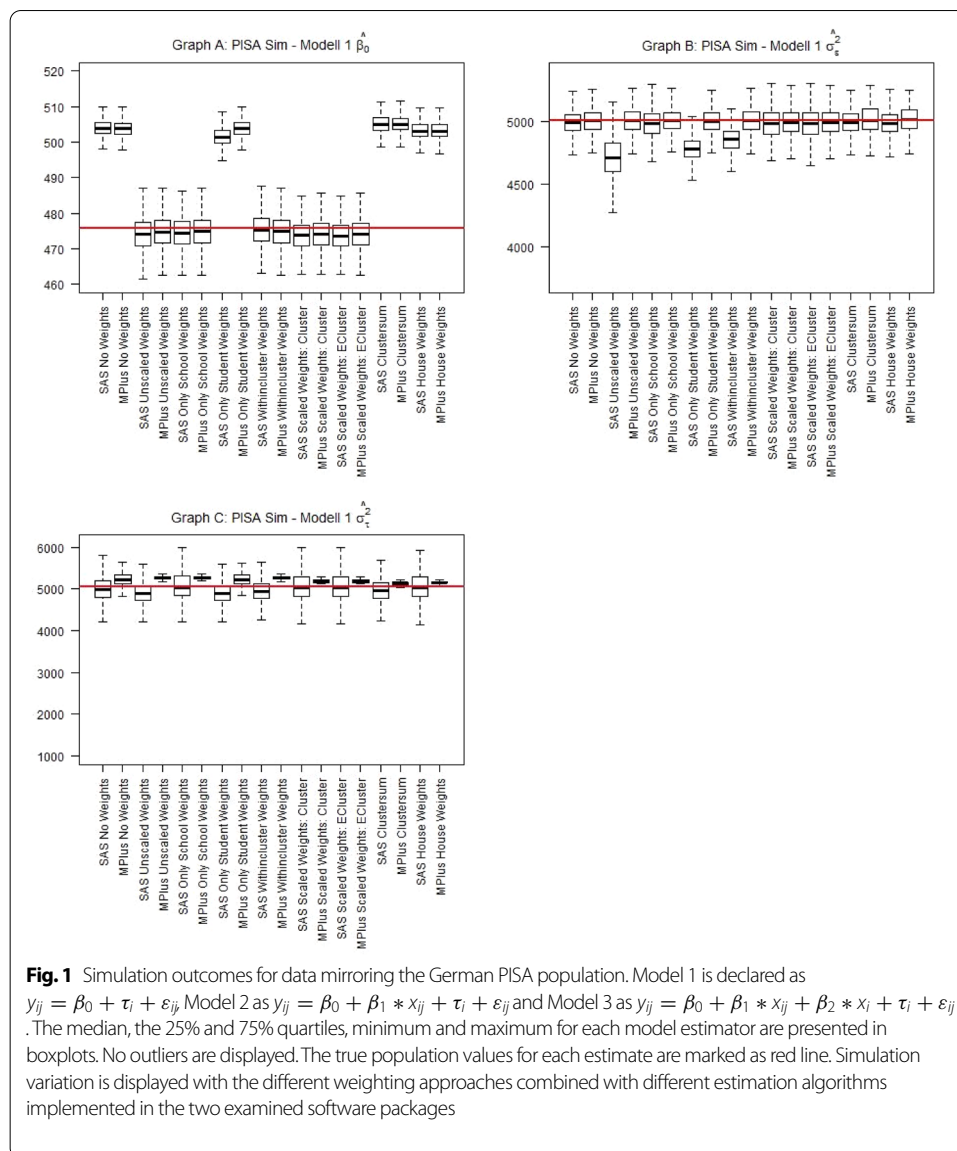
in Table 5 A, B and C. As the use of *Only School Weights* is more practical than using them plus scaling of the student weights (approaches *Cluster* and *Ecluster*), this approach would be the preferred one for both software programs SAS and Mplus, considering Simulation Scenario 1.

Outcomes for simulation scenario 2 (data reflecting low variances between schools)

Model 1

Having low variances between schools as simulated in Scenario 2, the estimated intercept distribution for $\hat{\beta}_0$ displayed in Fig. 2, Graph A, Graph D and Graph H, provides for all weighting approaches and both software program packages adequate estimators. Even the median seems to mask the true value, and interquartile spacing boxes do all overlap. As can also be seen in Table 6 A, B and C ($\hat{\beta}_0$) the coverage rates for all approaches are about or above 90%, which should preferably be higher, but are deemed acceptable in this study.

As in Simulation Scenario 1, the software program SAS again underestimates the *Variance Within* $\hat{\sigma}_\varepsilon^2$ applying the approaches *Unscaled Weights*, *Only Student Weights* and *Withincluster Weights* (Fig. 2, Graph B, Graph F and Graph K). This is verified in the low coverage rates between 0.3 and 0.8 from Table 6 A, B and C ($\hat{\sigma}_\varepsilon^2$). A different picture as in Simulation Scenario 1 can be seen for the estimation of the *Variance Between* $\hat{\sigma}_\tau^2$ in Simulation Scenario 2 (Fig. 2, Graph C, Graph G and Graph L). The *Variance Between* $\hat{\sigma}_\tau^2$ is incorrectly estimated by the approaches *Unscaled Weights*, *Only Student Weights* (only in Model 3, see Fig. 2, Graph L) and *Withincluster Weights*, in this case overestimated. It should be noted, however, that the

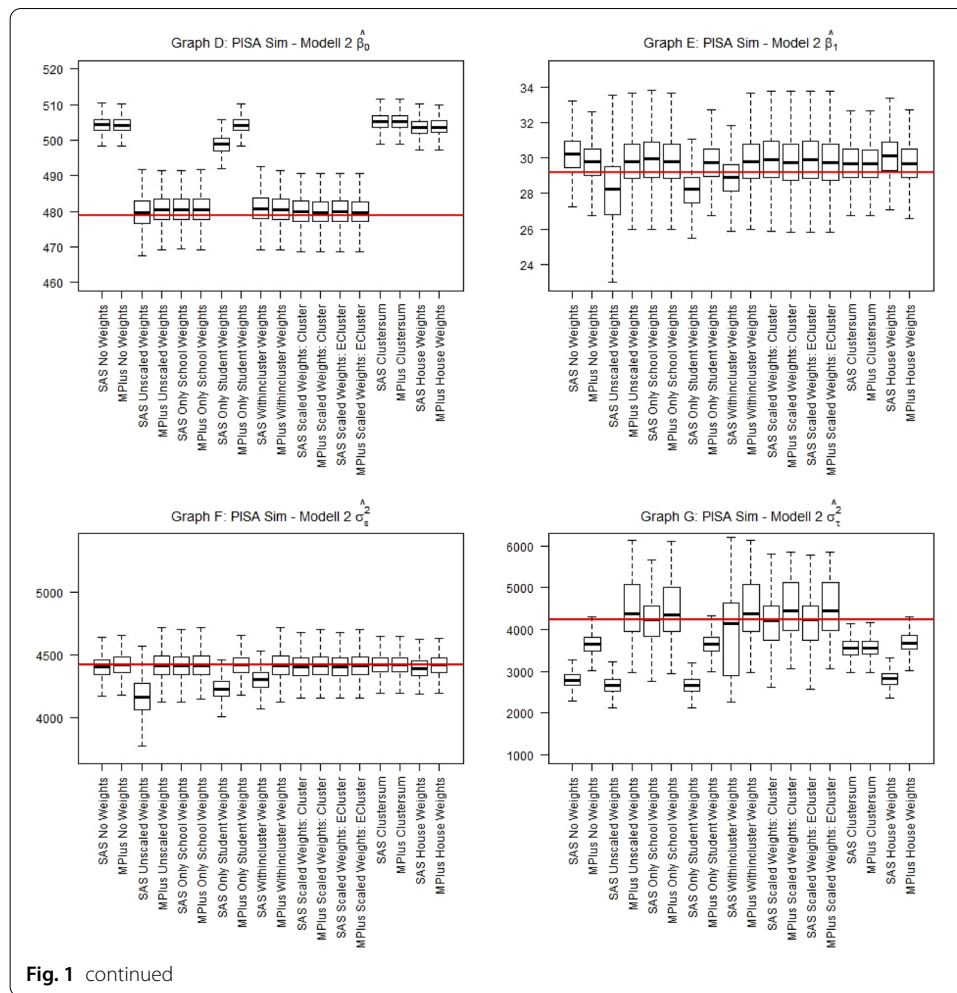


boxplots and whiskers do slightly overlap, which makes the statement to be interpreted with caution.

Models 2 and 3

Similar to Scenario 1, the estimation of the regressor $\hat{\beta}_1$ becomes more stable once this effect is added also at the school level; a finding confirmed by good coverage rates for both the regressor at student $\hat{\beta}_1$ and school level $\hat{\beta}_2$ in Table 6 C.

In Scenario 2, we also find that no distinctive difference between the two scaling techniques (*Cluster* and *Ecluster*) can be obtained, but the approach *Only School Weights* performs again equally well. Hence, as in Simulation Scenario 1, we would



recommend the weighting approach *Only School Weights* for both software program packages Mplus and SAS, respecting the specifications of Simulation Scenario 2.

Outcomes for simulation scenario 3 (data reflecting high variances between schools)

Models 1, 2 and 3

In the third considered scenario reflected in Fig. 3 (Simulation Scenario 3), we find nominal deviations from the two above described scenarios in the estimation of the *Variance Between* schools $\hat{\sigma}_{\tau}^2$. For Model 1 (Fig. 3, Graph C) and Model 3 (Fig. 3, Graph L) all weighting approaches provide correct estimates of this variance. Only in Model 2 (Fig. 2, Graph G) the *Variance Between* $\hat{\sigma}_{\tau}^2$ is underestimated by the software program SAS for all approaches, a finding being confirmed in very low coverage rates in Table 7 B. By adding the SES regressor at school level $\hat{\beta}_2$ into the model, this difference disappears and estimators of the *Variance Between* $\hat{\sigma}_{\tau}^2$ and the socio-economic background $\hat{\beta}_1$ and $\hat{\beta}_2$ become stable and unbiased. Also for Simulation Scenario 3 the weighting approach *Only School Weights* can be given as a clear recommendation for the use weighting in hierarchical models.

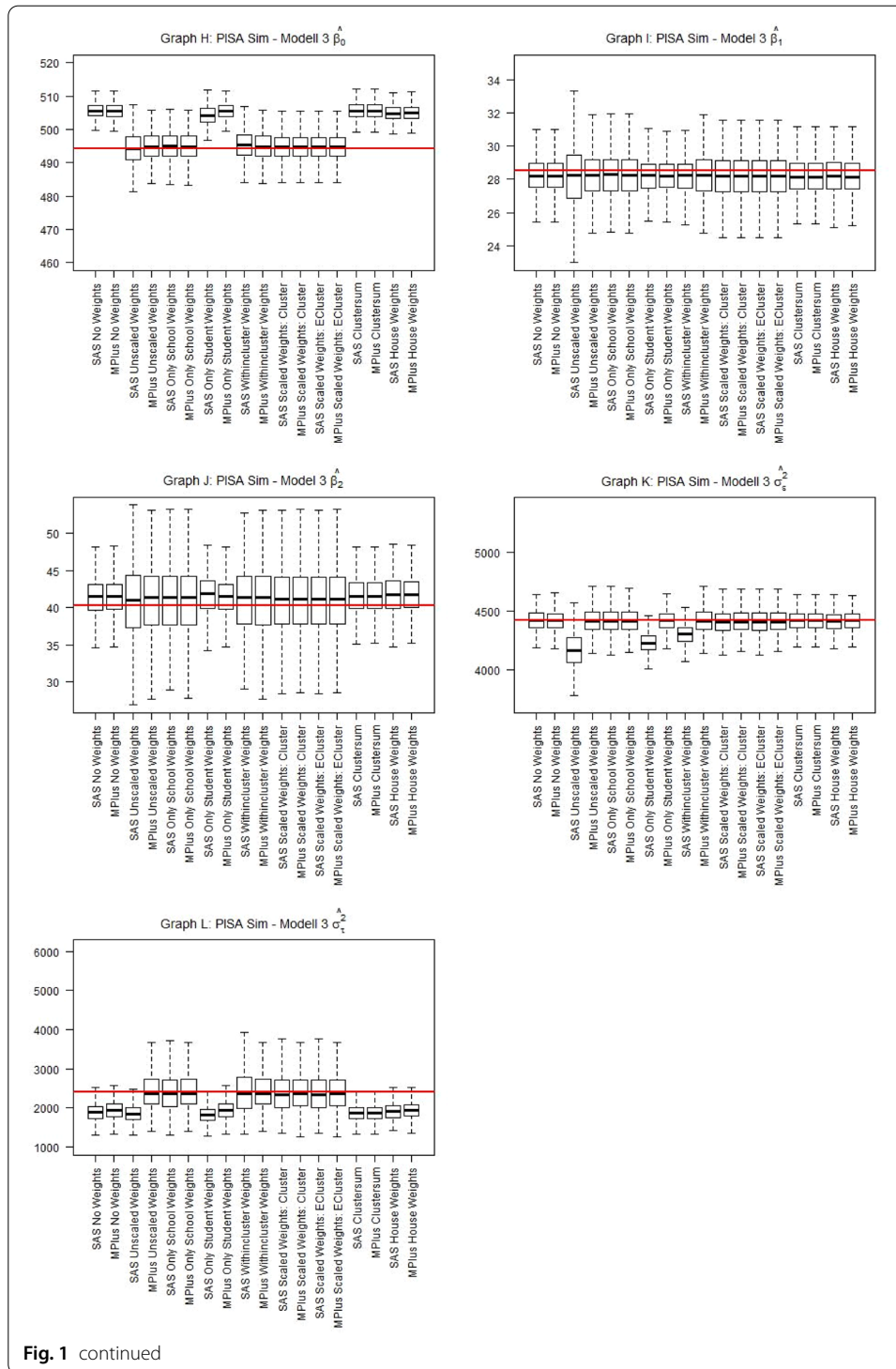
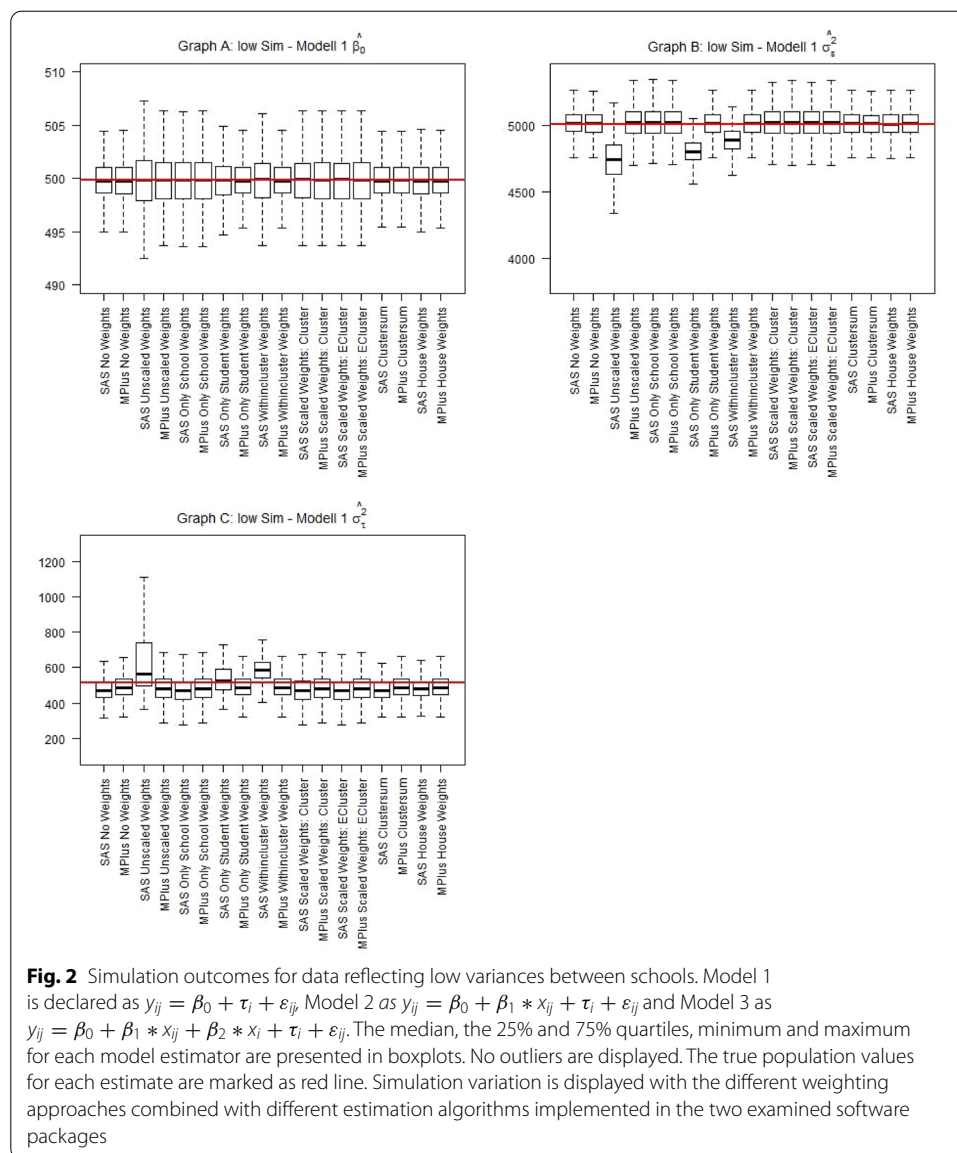


Fig. 1 continued

Software differences

Regarding the estimation accuracy of the software programs used, it can be said that Mplus provides slightly more precise estimates (e.g., Fig. 1, Graph I, or Table 5 B $\hat{\beta}_1$). Although the confidence intervals are sometimes quite small, they are partly more



biased (refer e.g., to Fig. 1, Graph L, or Table 5 $B \hat{\sigma}_{\tau}^2$). Like previously explained, this might be due to the different default settings like optimization algorithms in accelerating the EM algorithm. According to the SAS documentation and the analysis output, Quasi-Newton acceleration methods for optimization are used, whereas Mplus stated in their documentation to mainly use Quasi-Newton, but sometimes also other acceleration algorithms like Fisher-Scoring. The conditions under which to use one or the other method are not detailed. Instead, in the Mplus output, it is only declared that accelerating methods have been applied. Further, some algorithm starting default setting could also cause these differences. Beyond that, both software package declare to use pseudo ML estimation with the integration methods of adaptive quadrature. However, it must be clearly emphasized that with the recommended weighting approach using only schools weights, both software packages work equally well. If all

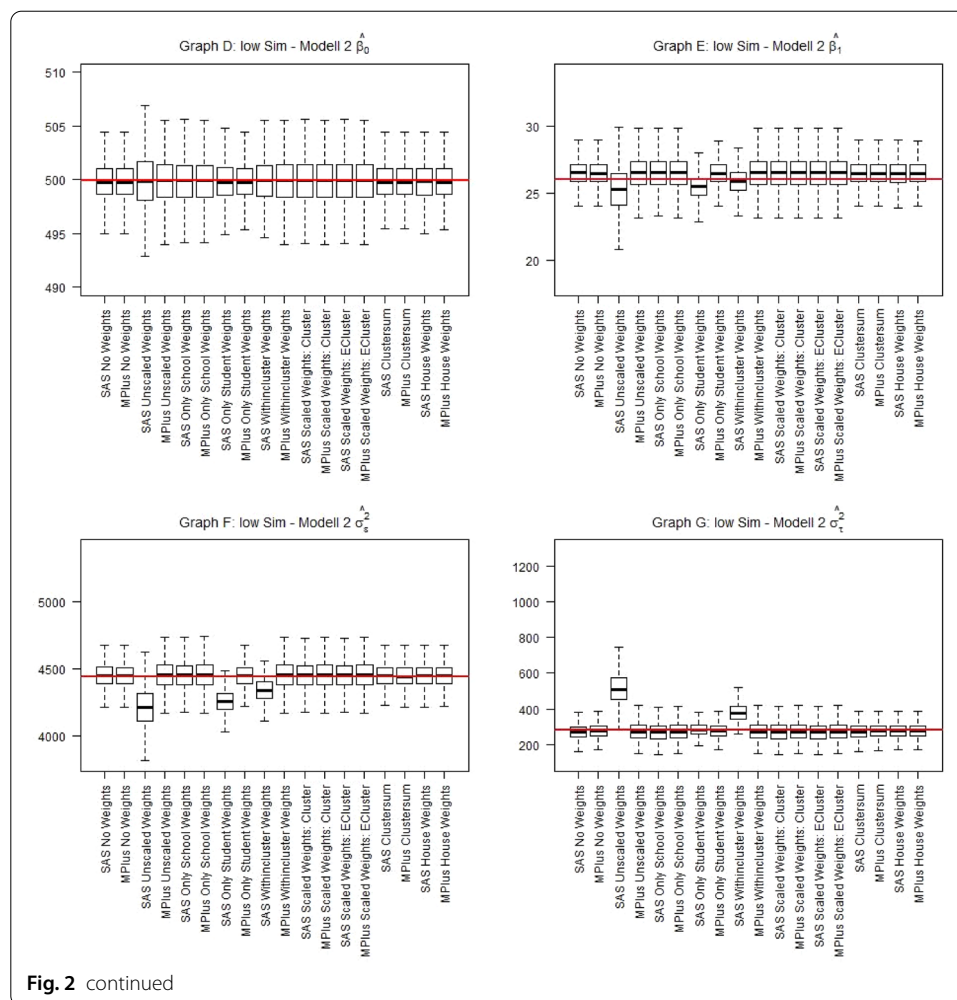


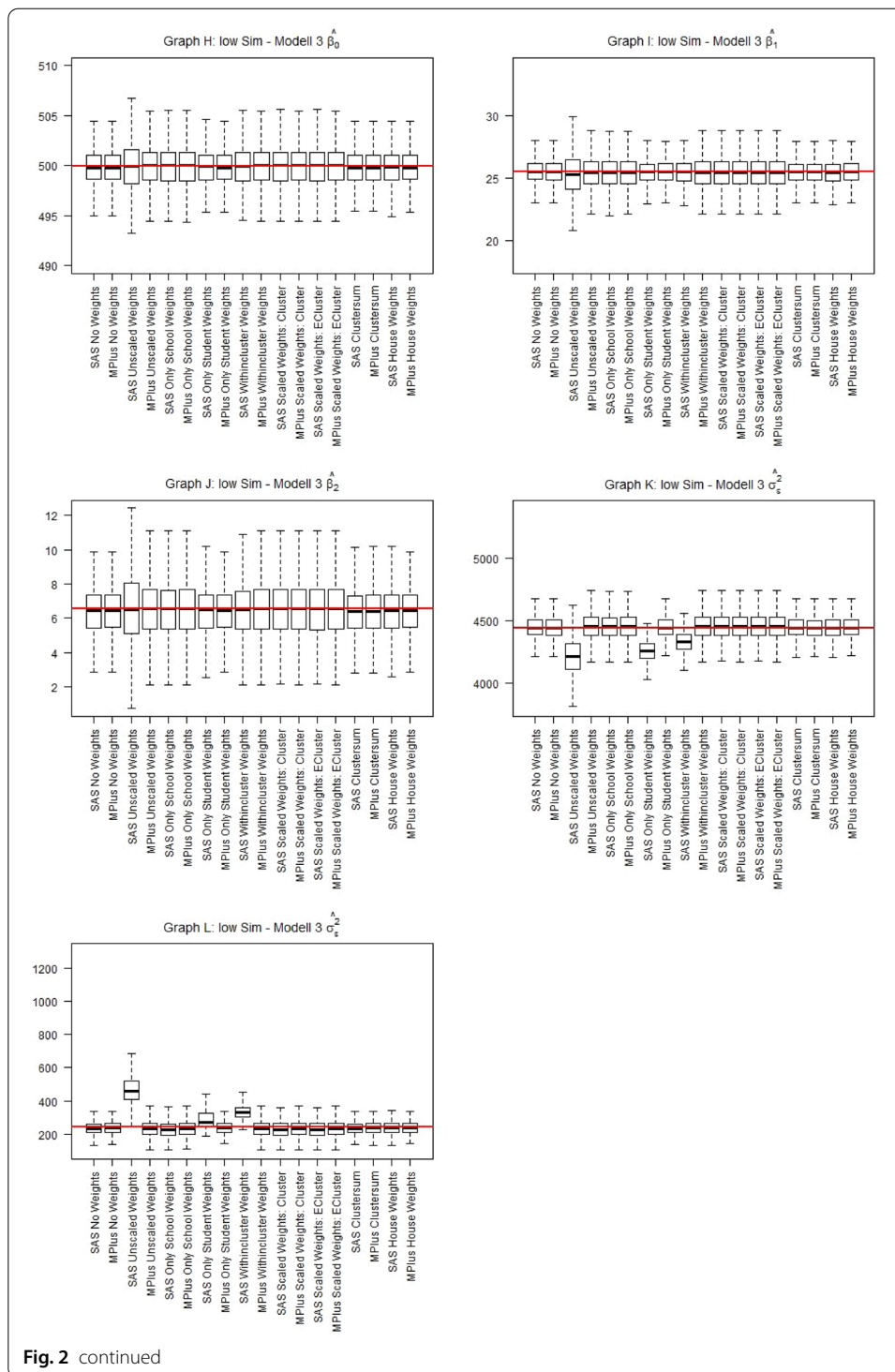
Fig. 2 continued

considerations for this and earlier scenarios are summarized, the authors recommend the weighting approach *Only School Weights* for all considered hierarchical models and scenarios for both software programs.

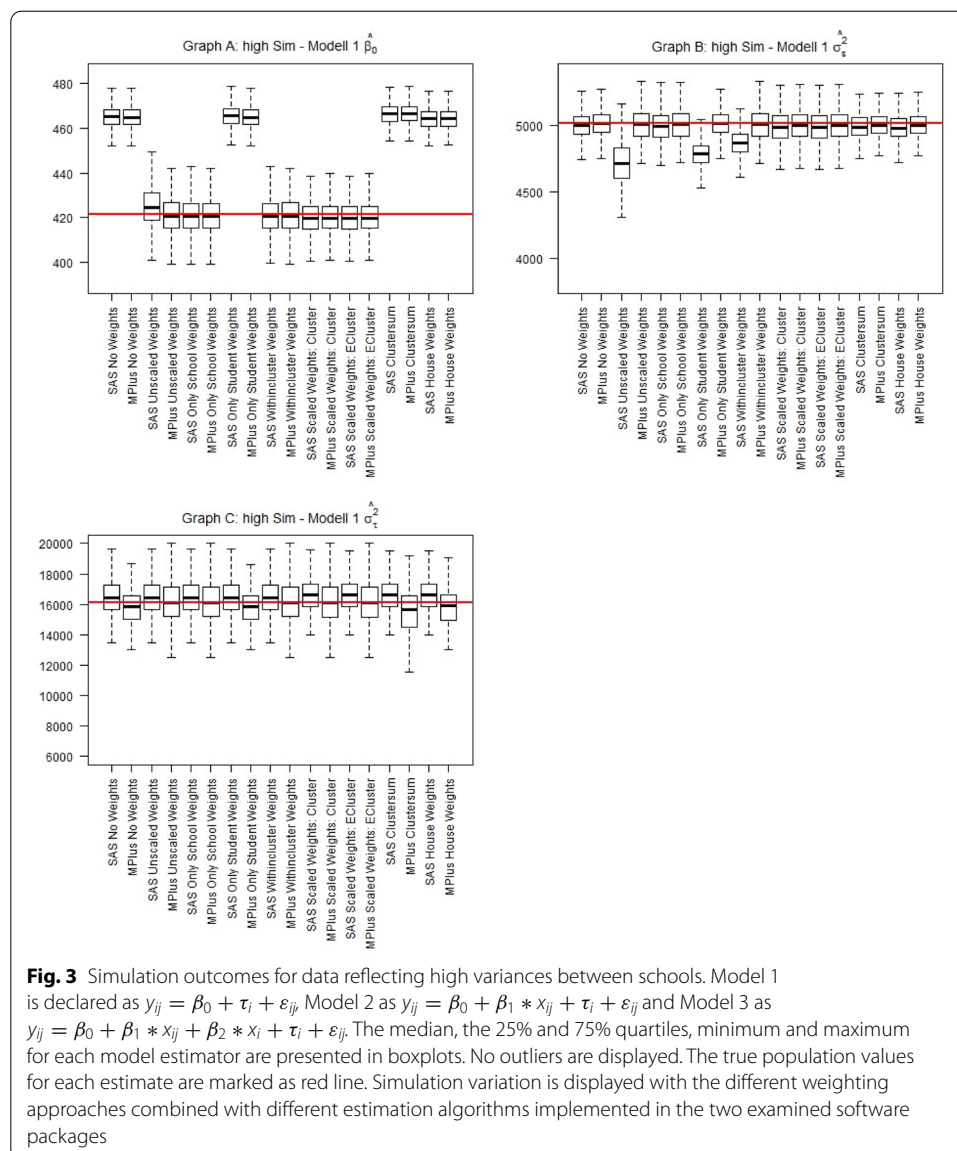
Application

In this section, we will apply our simulation results onto real data, covering a topic high on the research agenda in Germany and many other countries. With this, we would like to demonstrate the practical value of our study. We will first briefly look at previous publications in the field of multilevel analysis in connection with the PISA study, scientific literacy, and socio-economic background to show the significance of the topic. In a next step, we will apply multilevel regression models to the data from the PISA 2015 assessment (Reiss et al., 2018).

Germany is among the countries in the world where there remains a close relationship between socio-economic background and the performance of students, a fact which has been the cause of heavy public debate within the country. Using data from the PISA 2006 assessment, the OECD presented a hierarchical regression analysis



regarding the relationship between students' science competencies and the students' grade, the students' socio-economic background, the schools' socio-economic background, the students' migration background and the students' gender (OECD, 2007). For Germany, a higher science competence can be assumed for a higher grade and a higher



socio-economic background, for both the student and school level, whereas the school level (i.e., the average socio-economic background of students) has a higher impact on the results than students' personal socio-economic background. However, considering the findings presented earlier in this paper, we believe that the results must be interpreted with caution, as the weighting approach used for multilevel models in PISA 2006 (*House Weights*) did not show the best results in our simulation study. For the PISA 2015 cycle, the OECD (2016) reports a multilevel regression model with many factors related to the education systems, schools and students, again in connection to science literacy. They point out the positive (while negatively connoted) associations with science scores for both the OECD and all participating countries and economies. The OECD has changed its approach to weighting in multilevel models for this cycle, coinciding

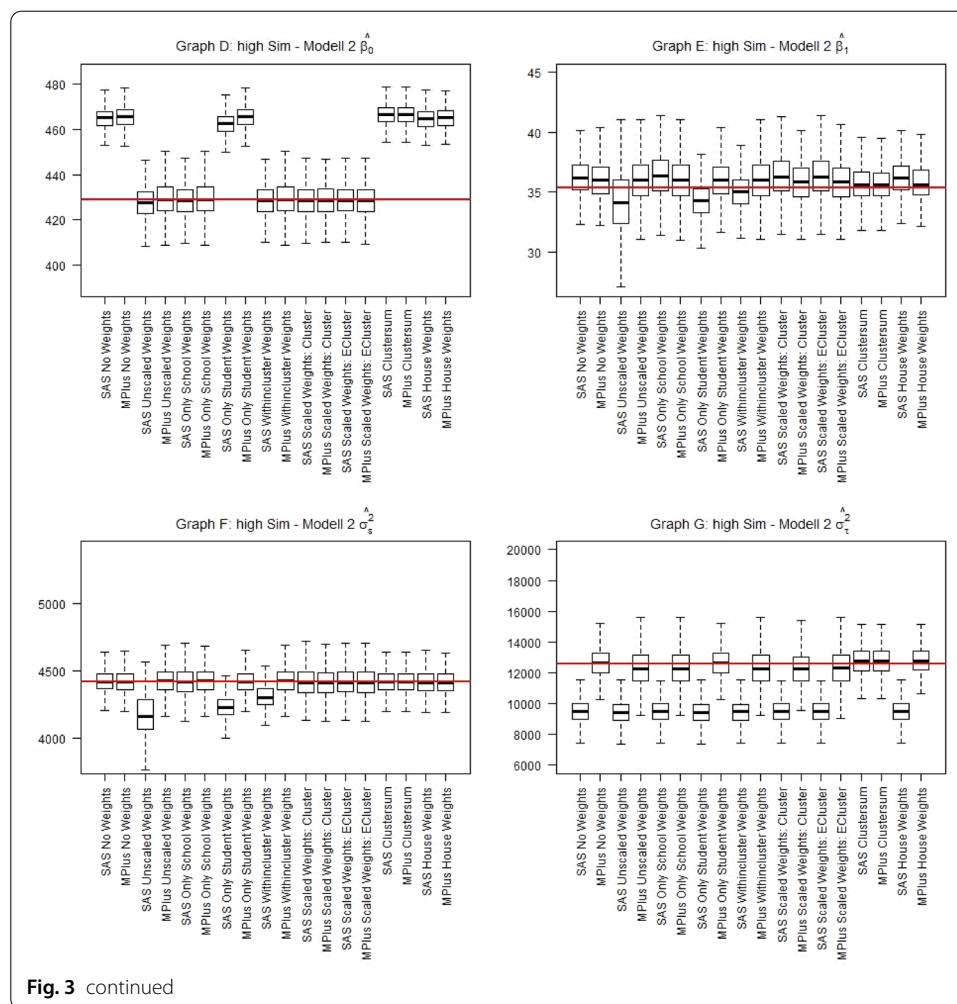


Fig. 3 continued

with the *Scaled Weights: Cluster* approach presented in this paper. Since this approach showed reliable results in our study, we believe these results can be trusted.

Apart from OECD publications, numerous papers have been published on the relationship of scientific literacy and socio-economic background. Papers relating to Asian countries stand out in particular. For example, Lam and Lau (2014) investigate how to improve science education in Hong Kong. Similarly, Sun et al. (2012) explore factors that affect students' science achievement in Hong Kong. Other publications are based on correlations between parents' attitudes towards science and the scientific competence of their children (Perera, 2014). Since the articles do not provide precise information on the exact use of the weights, these results should also be interpreted with caution.

Multilevel models uncovering factors at school and student level that determine students' performance, can offer significant and important evidence for policy makers. Obviously, they should be implemented in methodologically sound ways, which is why we present a practical application of the different weighting approaches studied in what follows.

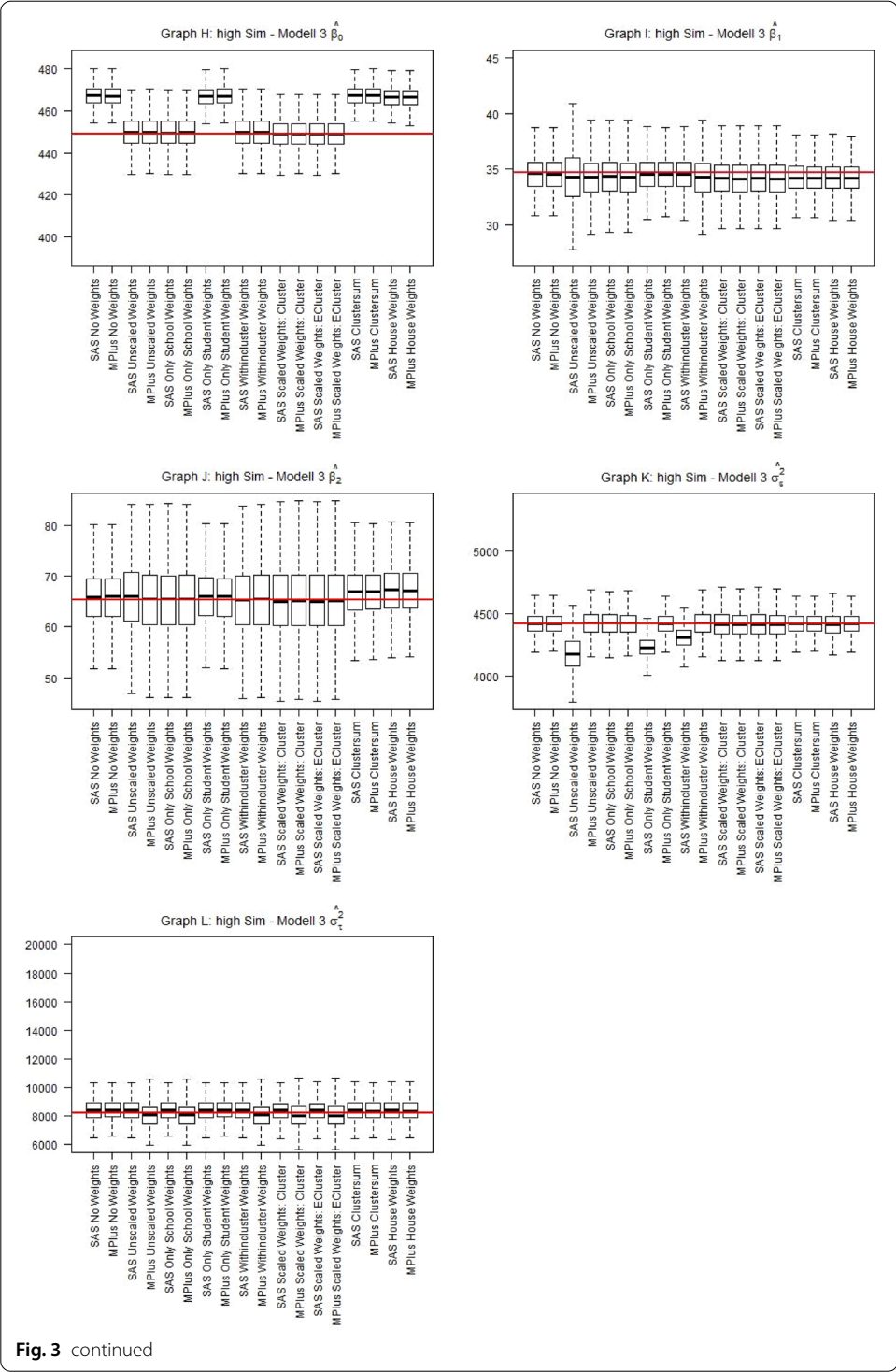


Table 8 Application PISA 2015 Data—Model 1 $y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$

| | $\hat{\beta}_0$ | SE $\hat{\beta}_0$ | $\hat{\sigma}_\varepsilon^2$ | SE $\hat{\sigma}_\varepsilon^2$ | $\hat{\sigma}_\tau^2$ | SE $\hat{\sigma}_\tau^2$ |
|--------------------------|-----------------|--------------------|------------------------------|---------------------------------|-----------------------|--------------------------|
| SAS | | | | | | |
| No weights | 508.28 | 4.49 | 5426.01 | 118.52 | 4013.66 | 272.93 |
| Unscaled weights | 481.80 | 5.41 | 5079.08 | 132.86 | 3940.63 | 284.36 |
| Only school weights | 483.83 | 5.37 | 5323.60 | 129.28 | 4042.81 | 331.98 |
| Only student weights | 506.42 | 4.54 | 5174.72 | 115.51 | 3936.82 | 236.50 |
| Withincluster weights | 484.43 | 5.34 | 5294.40 | 117.54 | 3987.60 | 307.94 |
| Scaled weights: cluster | 483.83 | 5.37 | 5323.60 | 129.28 | 4042.81 | 331.98 |
| Scaled weights: ECluster | 483.83 | 5.37 | 5323.60 | 129.28 | 4042.81 | 331.98 |
| Clustersum | 503.40 | 4.68 | 5451.93 | 121.11 | 4027.00 | 282.90 |
| House weights | 507.86 | 4.50 | 5421.13 | 121.32 | 4017.66 | 272.42 |
| Mplus | | | | | | |
| No weights | 507.92 | 4.50 | 5412.54 | 117.97 | 4811.59 | 372.69 |
| Unscaled weights | 483.67 | 5.34 | 5297.45 | 127.78 | 4865.75 | 454.33 |
| Only school weights | 483.57 | 5.37 | 5294.79 | 127.64 | 4860.81 | 456.38 |
| Only student weights | 507.81 | 4.50 | 5410.61 | 118.42 | 4818.42 | 373.28 |
| Withincluster weights | 483.57 | 5.37 | 5294.79 | 127.64 | 4860.81 | 456.38 |
| Scaled weights: cluster | 483.67 | 5.34 | 5297.45 | 127.78 | 4865.75 | 454.33 |
| Scaled Weights: ECluster | 483.69 | 5.34 | 5297.98 | 127.86 | 4863.82 | 454.18 |
| Clustersum | 504.26 | 4.61 | 5408.24 | 119.54 | 4760.76 | 373.38 |
| House weights | 507.92 | 4.50 | 5412.54 | 117.97 | 4811.59 | 372.69 |

Classifying the results of the simulation study to application data, the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages are displayed

Table 9 Application PISA 2015 Data – Model 2 $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$

| | $\hat{\beta}_0$ | SE $\hat{\beta}_0$ | $\hat{\beta}_1$ | SE $\hat{\beta}_1$ | $\hat{\sigma}_\varepsilon^2$ | SE $\hat{\sigma}_\varepsilon^2$ | $\hat{\sigma}_\tau^2$ | SE $\hat{\sigma}_\tau^2$ |
|--------------------------|-----------------|--------------------|-----------------|--------------------|------------------------------|---------------------------------|-----------------------|--------------------------|
| SAS | | | | | | | | |
| No weights | 509.57 | 4.06 | 15.87 | 1.21 | 5233.28 | 112.71 | 2670.31 | 181.69 |
| Unscaled weights | 484.76 | 5.14 | 11.31 | 1.54 | 4990.11 | 131.10 | 2239.82 | 143.37 |
| Only school weights | 487.12 | 5.02 | 13.62 | 1.37 | 5246.73 | 125.43 | 4149.08 | 414.42 |
| Only student weights | 507.36 | 4.20 | 13.12 | 1.20 | 5049.22 | 111.51 | 2229.14 | 125.10 |
| Withincluster weights | 487.64 | 4.96 | 13.96 | 1.19 | 5159.23 | 113.60 | 4219.66 | 404.56 |
| Scaled weights: cluster | 487.12 | 5.02 | 13.62 | 1.37 | 5246.73 | 125.43 | 4149.08 | 414.42 |
| Scaled weights: ECluster | 487.12 | 5.02 | 13.62 | 1.37 | 5246.73 | 125.43 | 4149.08 | 414.42 |
| Clustersum | 505.08 | 4.26 | 15.28 | 1.25 | 5300.05 | 116.11 | 3880.08 | 333.45 |
| House weights | 508.83 | 4.10 | 15.26 | 1.24 | 5293.84 | 117.33 | 3886.28 | 325.77 |
| Mplus | | | | | | | | |
| No weights | 509.19 | 4.10 | 15.10 | 1.20 | 5313.87 | 116.12 | 3876.26 | 324.59 |
| Unscaled weights | 487.20 | 4.99 | 13.60 | 1.37 | 5247.64 | 125.27 | 4141.49 | 411.45 |
| Only school weights | 487.11 | 5.01 | 13.61 | 1.37 | 5246.64 | 125.17 | 4151.67 | 414.03 |
| Only student weights | 509.15 | 4.10 | 15.12 | 1.21 | 5316.13 | 116.62 | 3877.56 | 325.67 |
| Withincluster weights | 487.11 | 5.01 | 13.61 | 1.37 | 5246.64 | 125.17 | 4151.67 | 414.03 |
| Scaled weights: cluster | 487.20 | 4.99 | 13.60 | 1.37 | 5247.64 | 125.27 | 4141.49 | 411.45 |
| Scaled weights: ECluster | 487.22 | 4.99 | 13.61 | 1.37 | 5248.73 | 125.38 | 4137.74 | 410.98 |
| Clustersum | 506.06 | 4.18 | 15.41 | 1.24 | 5304.56 | 116.65 | 3816.42 | 323.59 |
| House weights | 509.19 | 4.10 | 15.10 | 1.20 | 5313.87 | 116.12 | 3876.26 | 324.59 |

Classifying the results of the simulation study to application data, the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages are displayed

Table 10 Application PISA 2015 Data – Model 3 $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$

| | $\hat{\beta}_0$ | SE $\hat{\beta}_0$ | $\hat{\beta}_1$ | SE $\hat{\beta}_1$ | $\hat{\beta}_2$ | SE $\hat{\beta}_2$ | $\hat{\sigma}_\varepsilon^2$ | SE $\hat{\sigma}_\varepsilon^2$ | $\hat{\sigma}_\tau^2$ | SE $\hat{\sigma}_\tau^2$ |
|--------------------------|-----------------|--------------------|-----------------|--------------------|-----------------|--------------------|------------------------------|---------------------------------|-----------------------|--------------------------|
| SAS | | | | | | | | | | |
| No weights | 516.04 | 2.40 | 12.89 | 1.18 | 49.50 | 2.32 | 5316.47 | 116.13 | 1176.65 | 143.61 |
| Unscaled weights | 506.36 | 4.35 | 11.33 | 1.54 | 43.26 | 6.35 | 4972.29 | 130.45 | 1093.40 | 138.68 |
| Only school weights | 509.08 | 3.57 | 11.34 | 1.30 | 47.43 | 4.14 | 5256.11 | 126.86 | 1399.36 | 262.86 |
| Only student weights | 514.61 | 2.56 | 13.06 | 1.20 | 48.47 | 3.10 | 5049.66 | 111.53 | 1051.45 | 97.37 |
| Withincluster weights | 509.43 | 3.50 | 12.99 | 1.18 | 46.62 | 4.12 | 5170.39 | 114.15 | 1492.50 | 254.13 |
| Scaled weights: cluster | 509.08 | 3.57 | 11.34 | 1.30 | 47.43 | 4.14 | 5256.11 | 126.86 | 1399.36 | 262.86 |
| Scaled weights: ECluster | 509.08 | 3.57 | 11.34 | 1.30 | 47.43 | 4.14 | 5256.11 | 126.86 | 1399.36 | 262.86 |
| Clustersum | 515.11 | 2.52 | 13.02 | 1.21 | 48.04 | 2.78 | 5300.28 | 116.20 | 1218.51 | 166.33 |
| House weights | 515.81 | 2.42 | 13.07 | 1.20 | 48.79 | 2.53 | 5285.94 | 116.88 | 1222.41 | 156.36 |
| Mplus | | | | | | | | | | |
| No weights | 515.96 | 2.40 | 12.91 | 1.18 | 49.48 | 2.32 | 5320.55 | 116.31 | 1201.40 | 148.93 |
| Unscaled weights | 509.17 | 3.50 | 11.32 | 1.30 | 47.48 | 4.13 | 5271.26 | 127.35 | 1401.29 | 261.30 |
| Only school weights | 509.06 | 3.57 | 11.34 | 1.30 | 47.39 | 4.15 | 5270.65 | 127.31 | 1413.88 | 267.75 |
| Only student weights | 515.91 | 2.39 | 12.93 | 1.18 | 49.55 | 2.33 | 5322.62 | 116.78 | 1198.43 | 148.23 |
| Withincluster weights | 509.06 | 3.57 | 11.34 | 1.30 | 47.39 | 4.15 | 5270.65 | 127.31 | 1413.88 | 267.75 |
| Scaled weights: cluster | 509.17 | 3.50 | 11.32 | 1.30 | 47.48 | 4.13 | 5271.26 | 127.35 | 1401.29 | 261.30 |
| Scaled weights: ECluster | 509.19 | 3.50 | 11.32 | 1.30 | 47.49 | 4.12 | 5272.64 | 127.49 | 1396.79 | 259.89 |
| Clustersum | 515.41 | 2.46 | 13.13 | 1.20 | 47.86 | 2.79 | 5315.53 | 117.16 | 1210.36 | 160.83 |
| House weights | 515.96 | 2.40 | 12.91 | 1.18 | 49.48 | 2.32 | 5320.55 | 116.31 | 1201.40 | 148.93 |

Classifying the results of the simulation study to application data, the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages are displayed

For the analyses with PISA 2015 data, the same three hierarchical models as applied in the simulation study were used. The first plausible value (PV) for the domain of Science approximates the distribution of student achievement correctly (Davies et al., 2009). The socio-economic background is represented by the z-standardized variable ESCS for both the school and student level. As in the simulation study, the variance within and between schools will be estimated. The same weighting approaches as in the simulation study are also investigated here. The different results can be correctly classified using the results of the simulation study. Therefore, the weighting approach *Only School Weights* is assumed in the following as a reference point for the recommended implementation of the weights and thus as the correct interpretation approach for the explanation of the estimated parameters of the hierarchical models. Both estimation methods represented in the different software packages are used to get a more comprehensive picture of the application.

Tables 8, 9, 10 show the different results for Models 1, 2 and 3, each displayed for both software packages, respectively.

The weighting scenarios *No Weights*, *Only Student Weights*, *Clustersum* and *House Weights* achieve higher values for the intercept $\hat{\beta}_0$ than the approach *Only School Weights*. This applies to both software packages used and all defined models. These values are estimated too highly and can lead to a misinterpretation of the intercept as a value too high for the mean of the schools' achievement. Concerning the link between scientific achievement and socio-economic background, it can be stated that with regard to the reference method *Only School Weights*, all weighting approaches and software

packages estimated this context correctly for both the student and the school level. Having a higher socio-economic background, a higher achievement for the students is estimated. This correlation is even more pronounced for the socio-economic background at school level.

Regarding the *Variance Between* the schools $\widehat{\sigma}_{\tau}^2$, it can be noted that compared to the reference approach *Only School Weights* this variance is underestimated for Models 2 and 3 (Tables 9 and 10), for both software packages and for the weighting scenarios *No Weights*, *Unscaled Weights* (only SAS), *Only Student Weights*, *Clustersum* and *House Weights*. This underestimation may result in less variability being assumed between schools than is actually present in the population. Also, with regard to the *Variance Within* $\widehat{\sigma}_{\varepsilon}^2$, caution is advised in connection with the weighting variants *Unscaled Weights*, *Only Student Weights* and *Withincluster Weights*. Compared to the approach *Only School Weights*, these variances are also underestimated with the software SAS.

In summary, with the help of the simulation study, the application of PISA data demonstrated that the influence of school-specific aspects on student performance is of great importance and therefore a consideration of the hierarchies in PISA analyses, using the best-performing estimation approach, is highly recommended.

Summary and conclusions

In order to determine the best weighting scheme in hierarchical models with LSA data, a simulation study based on the PISA data structure was performed examining different weighting approaches and scaling techniques frequently used in the research community. Further, two different software packages, Mplus (with default two-level analysis settings) and SAS (with its procedure PROC GLIMMIX), were compared against each other with a focus on deployed estimation procedures and algorithms. In summary, this study provides a comprehensive picture of many possible and previously used weighting approaches. This research program implies which weighting approach leads to the most precise and least biased estimation of parameters in multilevel models with LSA data, and thus gives clear guidance which approach should be used for such analysis.

We were able to show that the weighting scenarios *Only School Weights*, *Scaled Weights: Cluster* and *Scaled Weights: ECluster* provide the least biased and sufficiently precise parameter estimates throughout all three considered models, and in all three simulation scenarios. As the use of *Only School Weights* is easier to implement than the other well-performing methods, we recommend this approach, independently of whether SAS or Mplus is being used.

It can be noted that the software program SAS with its used procedure PROC GLIMMIX, provides larger quartile spacing's or more wrongly estimated variances than the software package Mplus with its used default settings for two-level analysis. As both software packages SAS and Mplus are not as transparent as freely available software packages like R (R Core Team, 2018), we can only assume where the distinction between the software programs are, although the authors have put a lot of time and effort into finding internal settings of these programs. Although both software packages provide quite good consulting services, they lack insight into the actual internal procedures of the syntaxes used.

Furthermore, no explicit difference was found comparing the considered scaling techniques of level one weights. The scaling technique resulting in student weights summing up to the cluster size as well as the technique where student weights sum up to the effective sample size within clusters, perform nearly the same for all simulation scenarios and analysed models. Therefore, both methods seem to be legitimate. Nevertheless, the authors would like to reiterate the importance of applying school weights at level two, as they have significant effects on most parameter estimates, and seem to be needed to sufficiently reflect the LSA sample design in multilevel models, as it is characterized by significantly varying school selection probabilities. Level one weights may not be as important, because the student weights have by design a low variety within schools.

Applying the investigated weighting scenarios to real PISA data, we could show the potential threads on validity of results and interpretation when using different weighting methods than the recommended ones.

Limiting the explanatory power of this study is the number of relatively simple models considered. Further research is needed to evaluate the findings for more advanced hierarchical models; for example, with random slopes, or those including multiple predictor variables, all introducing further error terms. In particular, immigration background, student gender and the type of school attended, for example, are also potential predictors of the relationship between competence and social background. Finally, other frequently used software programs like HLM (Raudenbush, 2007) could also be examined.

Implications for practice

This simulation study has shown that using only the school weights provide the most unbiased estimates for hierarchical models. In this approach, the final school weights are specified as level two weights, while no weight is used at level one. Final school weights reflect the school selection probabilities, adjusted for school nonresponse, and are typically provided with the public datasets of LSA. For PISA data, the respective variable is named *nonresponse adjusted school base weight* *W_NRASCHBWT* in former PISA cycles, e.g. OECD (2017). Hence, the identified preferable HLM weighting method is at the same time one that can be implemented in a straightforward manner. This weighting approach may be useful as well for other LSA with a similar data structure, i.e., individuals nested within clusters. Such data are for example student and teacher data of ICILS, and teacher data of ICCS. Within some limits the findings are even applicable to data with slightly different structure, e.g., with class sampling such as TIMSS, PIRLS and ICCS student data. For the latter datasets, the school weight variable is called “Final school weight”—users are referred to the technical documentation of the studies for the respective variable names. We are confident that the findings can even be generalized to other data with similar hierarchical structure outside the education sector, that is, data coming from two-stage samples with varying selection probabilities at stage one, but uniform selection probabilities at stage two. Regarding the investigated software packages Mplus and SAS, no significant differences between the programs become visible with the preferred weighting approach of using only final school weights.

We are confident that the recommended weighting approach will help many researchers in the application of MLM with weights, thus driving further insightful research in the field of LSA.

Abbreviations

CR: Coverage rate; EM: Expectation–maximisation; ESCS: Economic, Social and Cultural Index; HLM: Hierarchical linear modelling; HT: Horwitz–Thompson estimator; ICC: Intraclass correlation; ICCS: International Civic and Citizenship Education Study; ICILS: International Computer and Information Literacy Study; IGLS: Iterative generalised least squares; LSA: Large-scale assessment; ML: Maximum likelihood; MLM: Multilevel modelling; PIRLS: Progress in International Reading Literacy Study; PISA: Programme for International Student Assessment; PML: Pseudo maximum likelihood; PPS: Probability proportional to size; PWGLS: Probability weighted generalized least squares; PV: Plausible values; SE: Standard error; TIMSS: Trends in International Mathematics and Science Study.

Acknowledgements

Not applicable.

Authors' contributions

All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹ TUM School of Education, Centre for International Student Assessment (ZIB), Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany. ² Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany. ³ International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany. ⁴ Centre for Teacher Education, University of Vienna, Vienna, Austria.

Received: 10 August 2020 Accepted: 17 March 2021

Published online: 26 March 2021

References

- Asparouhov, T. (2004). Weighting for unequal probability of selection in multilevel modeling. *Mplus Web Notes*: No. 8.
- Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics Theory and Methods*, 35(3), 439–460. <https://doi.org/10.1080/03610920500476598>.
- Bertolet, M. (2008). *To Weight or not to weight? Incorporating sampling designs into model-based analyses*. Dissertation. Pittsburgh: Carnegie Mellon University.
- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review / Revue Internationale De Statistique*, 51(3), 279–292. <https://doi.org/10.2307/1402588>.
- Binder, D. A., & Roberts, G. (2010). Design- and model-based inference for model parameters. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics*. (Vol. 29, pp. 33–54). Elsevier North-Holland. [https://doi.org/10.1016/S0169-7161\(09\)00224-7](https://doi.org/10.1016/S0169-7161(09)00224-7).
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443–459. <https://doi.org/10.1007/BF02293801>.
- Brewer, K. R. W., & Hanif, M. (1983). *Sampling with unequal probabilities lecture notes in statistics*. (Vol. 15). Springer. <https://doi.org/10.1007/978-1-4684-9407-5>.
- Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. *Sociological Methodology*, 43(1), 178–219. <https://doi.org/10.1177/0081175012460221>.
- Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, 9, 49. <https://doi.org/10.1186/1471-2288-9-49>.
- Chambers, J. M. (1983). *Graphical methods for data analysis*. Chapman & Hall statistics series. Wadsworth & Brooks/Cole.
- Chantala, K., Blanchette, D., & Suchidnran, C. (2011). *Software programs to compute sampling weights for multilevel analysis*. University of North Carolina at Chapel Hill.
- Chantala, K., & Suchidnran, C. (2006). Adjusting for unequal selection probability in multilevel models: A comparison of software packages. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 2815–2824.

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 39(1), 1–38. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>.
- Gebhardt, E., Ainley, J., Fraillon, J., Friedman, T., & Schulz, W. (2014). *Preparing for life in a digital age: The IEA International Computer and Information Literacy Study International Report*. International Association for Educational Achievement (IEA).
- Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73(1), 43–56. <https://doi.org/10.1093/biomet/73.1.43>.
- Graubard, B. I., & Korn, E. L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, 5(3), 263–281. <https://doi.org/10.1177/096228029600500304>.
- Grilli, L., & Pratesi, M. (Eds.). (2005). *Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs*. Wiley-InterScience. <https://doi.org/10.1002/0471667196>.
- Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685. <https://doi.org/10.1080/01621459.1952.10483446>.
- Jamshidian, M., & Jennrich, R. I. (1997). Acceleration of the EM algorithm by using quasi-newton methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3), 569–587. <https://doi.org/10.1111/1467-9868.00083>.
- Kish, L. (1965). *Survey sampling*. Wiley.
- Korn, E. L., & Graubard, B. I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 175–190. <https://doi.org/10.1111/1467-9868.00379>.
- Lai, M. H. C., & Kwok, O. (2015). Examining the rule of thumb of not using multilevel modeling: The “Design Effect Smaller Than Two” rule. *The Journal of Experimental Education*, 83(3), 423–438. <https://doi.org/10.1080/00220973.2014.907229>.
- Lam, T. Y. P., & Lau, K. C. (2014). Examining factors affecting science achievement of Hong Kong in PISA 2006 using hierarchical linear modeling. *International Journal of Science Education*, 36(15), 2463–2480. <https://doi.org/10.1080/09500693.2013.879223>.
- Lange, K. (1995). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, 5(1), 1–18.
- Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4), 817–827. <https://doi.org/10.1093/biomet/74.4.817>.
- Martin, M. O., & Mullis, I. (2013). *Timss and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—implications for early learning*. International Association for the Evaluation of Educational Achievement (IEA).
- Meinck, S. (2020). Sampling, weighting, and variance estimation. In H. Wagemaker (Ed.), *IEA research for education: v 10 Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement*. (pp. 113–129). Springer. https://doi.org/10.1007/978-3-030-53081-5_7.
- Meinck, S., & Vandenplas, C. (2012). Sample size requirements in HLM: An empirical study. IER Institute, IERI monograph series issues and methodologies in large-scale assessments. *Special Issue 1, Educational Testing Service and International Association for the Evaluation of Educational Achievement*.
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247), 335. <https://doi.org/10.2307/2280232>.
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide: Eighth Edition*.
- OECD. (2007). *Pisa 2006, science competencies for tomorrow's world*. (Vol. I). OECD Publishing. <https://doi.org/10.1787/9789264040014-en>.
- OECD. (2009). *Pisa data analysis manual: Spss*. (2nd ed.). OECD Publishing. <https://doi.org/10.1787/9789264056275-en>.
- OECD. (2011). *PISA 2009 results: Students on line: Digital technologies and performance*. (Vol. VI). OECD Publishing. <https://doi.org/10.1787/9789264112995-en>.
- OECD. (2014). *Pisa 2012 Results: Excellence through Equity Giving every student the chance to succeed*. (Vol. 2). OECD Publishing. <https://doi.org/10.1787/9789264201132-en>.
- OECD. (2016). *Policies and practices for successful schools/PISA 2015 Results. PISA 2015 results*. (Vol. II). OECD Publishing. <https://doi.org/10.1787/9789264267510-en>.
- OECD. (2017). *PISA 2015 technical report*. OECD Publishing.
- OECD. (2019). *PISA 2018 results what school life means for students' lives*. (Vol. III). OECD Publishing. <https://doi.org/10.1787/acd78851-en>.
- Perera, L. D. H. (2014). Parents' attitudes towards science and their children's science achievement. *International Journal of Science Education*, 36(18), 3021–3041. <https://doi.org/10.1080/09500693.2014.949900>.
- Petkova, M. (2016). *Using sampling weights in multilevel analysis of PISA data* (Master Thesis). Ludwig-Maximilians-Universität München.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review / Revue Internationale De Statistique*, 61(2), 317. <https://doi.org/10.2307/1403631>.
- Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5(3), 239–261. <https://doi.org/10.1177/096228029600500303>.
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1), 23–40. <https://doi.org/10.1111/1467-9868.00106>.
- Pfeffermann, D., & Sverchkov, M. (2010). Inference under informative sampling. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics*. (Vol. 29, pp. 455–487). Elsevier North-Holland. [https://doi.org/10.1016/S0169-7161\(09\)00239-9](https://doi.org/10.1016/S0169-7161(09)00239-9).
- R Core Team. (2018). *R: A language and environment for statistical computing*. [Computer software]. R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4), 805–827. <https://doi.org/10.1111/j.1467-985X.2006.00426.x>.

- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal: Promoting Communications on Statistics and Stata*, 2(1), 1–21. <https://doi.org/10.1177/1536867X0200200101>.
- 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. <https://doi.org/10.1016/j.jeconom.2004.08.017>.
- Raudenbush, S. (2007). *Hlm 6: Hierarchical linear and nonlinear modeling*. . Scientific Software International.
- Reiss, K., Sälzer, C., Schiepe-Tiska, A., Mang, J., Heine, J.-H., Weis, M., . . . Köller, O. (2018). *Programme for International Student Assessment 2015 (PISA 2015): Dataset*. https://doi.org/10.5159/IQB_PISA_2015_v1
- RStudio Team. (2018). RStudio: Integrated Development Environment for R (Version 1.1.456) [Computer software]. RStudio, Inc. Retrieved from <http://www.rstudio.com/>
- Rust, K. F., & Rao, J. N. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5(3), 283–310. <https://doi.org/10.1177/096228029600500305>.
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142–151. <https://doi.org/10.3102/0013189X10363170>.
- Särndal, C.-E., Swensson, B., & Wretman, J. H. (2003). *Model assisted survey sampling*. Springer series in statistics. . Springer.
- SAS Institute Inc. (2018). SAS-STAT Software (Version 9.4) [Computer software]. Cary, NC. Retrieved from <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*. International Association for Educational Achievement (IEA). <https://doi.org/10.1007/978-3-319-73963-2>
- Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, & T. Smith (Eds.), *Wiley series in probability and mathematical statistics: Applied probability and statistics. Analysis of complex surveys*. (pp. 59–88). Wiley.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. . SAGE Publishing.
- Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4), 475–502. https://doi.org/10.1207/S15328007SEM0904_2.
- Sun, L., Bradley, K. D., & 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. <https://doi.org/10.1080/09500693.2012.708063>.
- von Davier, M., Gonzalez, E., & Myslevy, R. (2009). What are plausible values and why are they useful?: IER Institute, IERI monograph series issues and methodologies in large-scale assessments. *Special issue 2, educational testing service and international association for the evaluation of educational achievement*.
- West, B. T., & Galecki, A. T. (2012). An overview of current software procedures for fitting linear mixed models. *The American Statistician*, 65(4), 274–282. <https://doi.org/10.1198/tas.2011.11077>.
- Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). The box plot: A simple visual method to interpret data. *Annals of Internal Medicine*, 110(11), 916–921. <https://doi.org/10.7326/0003-4819-110-11-916>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.