Evaluating the PISA sampling design by stratification schemes and weighting procedures in hierarchical modelling

Dissertation von Julia Mang

München 2023

Evaluating the PISA sampling design by stratification schemes and weighting procedures in hierarchical modelling

Dissertation an der Fakultät für Mathematik, Informatik und Statistik

der Ludwig-Maximilians-Universität München

eingereicht von

Julia Mang

am 02.08.2023

Erstgutachter: Prof. Dr. Helmut Küchenhoff

Zweitgutachter: Prof. Dr. Manfred Prenzel

Externe Gutachterin: Prof. Dr. Leslie Rutkowski

Tag der mündlichen Prüfung: 13.11.2023

Summary

The *Programme for International Student Assessment* (PISA) is an important assessment tool. As a worldwide monitoring study of basic educational competencies of fifteen-year-old students, it allows to conclude education systems, long-term educational investments and identify educational development and referred changes in time. In PISA, a state-of-the-art sampling design acknowledged by the scientific community is applied (Rutkowski et al., 2013). As this complex sampling design must be accounted for in the study's analyses, statistical techniques, and procedures were developed. To evaluate improving alternatives in the complexity of these methods, it is essential to constantly conduct theoretical considerations and associated simulation studies (Boulesteix et al., 2020). In this dissertation, two PISA-sampling-related topics were examined in detail. New derived suggestions for substantial improvements to the PISA sampling design were built based on these findings. Both studies presented in this dissertation look at sampling-related concepts (both model groups of sampling units), i.e., weighting in hierarchical linear models and stratification in the PISA sampling process. Those concepts must be correctly represented in analyses to avoid biased standard error estimators.

In the first study, we determine under theoretical consideration and simulation which stratification scheme is best for PISA in Germany. Thus, we examine seven different stratification designs – selected according to scenarios used in past large-scale assessment studies in Germany – and theoretical, new devised approaches for future implementations. As a result of this examination, we recommend a stratification of grouped German federal states and designs using school types as explicit and federal states as implicit stratifiers.

In the second study, we identify the best utilisation of sampling weights in hierarchical linear modelling based on theoretical considerations and simulative results. We examine nine different weighting designs. The selected sampling scenarios are based on framing approaches to explain required weighting in hierarchical modelling, settings promoted in the literature and theoretical, new devised considerations for future implementations. We consider different estimation, optimization, acceleration methods, and approaches to using sampling weights. The results reveal three weighting approaches performing best in retrieving the true population parameters. One implies using only level two weights (here: final school weights). Due to its simple implementation, it is the most favorable one.

Zusammenfassung

Das *Programme for International Student Assessment* (PISA) ist als weltweite Beobachtungs-Studie der Grundkompetenzen 15-jähriger Schüler:innen ein sehr wichtiges Bewertungsinstrument, um Schlüsse über Bildungssysteme und langfristige Bildungsinvestitionen zu ziehen sowie um Bildungsentwicklungen im Zeitverlauf zu ermitteln. Bei PISA wird ein wissenschaftlich fundiertes und anerkanntes Stichprobendesign angewandt (Rutkowski et al., 2013). Da dieses komplexe Stichprobendesign auch in den Auswertungsmethoden der Studie berücksichtigt werden muss, wurden vielfach zitierte Analysetechniken und -verfahren entwickelt. Um diese komplexen Methoden regelmäßig zu überprüfen und gegebenenfalls zu verbessern, ist es wichtig, wiederholt theoretische Weiterentwicklungen und entsprechende Überprüfung durch Simulationsstudien durchzuführen (Boulesteix et al., 2020). In dieser Dissertation werden zwei PISA-Stichprobenverfahren näher untersucht. Auf der Grundlage dieser Erkenntnisse werden neue Vorschläge für wesentliche Verbesserungen des PISA-Stichprobendesigns abgeleitet. Beide Studien befassen sich mit stichprobenbezogenen Konzepten. Im Speziellen sind dies Gewichtungen in hierarchischen linearen Modellen und Stratifizierungsverfahren der PISA-Stichprobe. Es ist von großer Bedeutung, dass sie in den Analysen korrekt dargestellt werden, um verzerrte Standardfehlerschätzer zu vermeiden.

In der ersten Studie untersuchen wir Theorie geleitet und mit einem simulativen Ansatz, welches Stratifikationsschema für PISA in Deutschland am besten ist. Dazu untersuchen wir sieben verschiedene Stratifikationsdesigns – ausgewählt auf der Grundlage von Szenarien, welche in vergangenen large-scale assessment Studien in Deutschland verwendet wurden – sowie theoretische und neu entwickelte Überlegungen für zukünftige Implementierungen. Als Ergebnis dieser Studie empfehlen wir eine Stratifizierung von gruppierten Bundesländern sowie Szenarien, die lediglich Schulformen als explizite und zusätzlich Bundesländer als implizite Stratifizierung verwenden.

In der zweiten Studie ermitteln wir auf Grundlage theoretischer Überlegungen und simulativer Ergebnisse die beste Anwendung von Stichprobengewichte in hierarchischen linearen Modellen. Wir betrachten neun verschiedene Ansätze zur Verwendung dieser Gewichte. Die ausgewählten Stichprobenszenarien basieren auf Rahmenansätzen zur Erklärung der erforderlichen Gewichtung in der hierarchischen Modellierung, auf in der Literatur zitierten Verfahren und auf theoretischen, neu entwickelten Überlegungen für zukünftige Implementierungen. Wir betrachten verschiedene Schätzmethoden, inklusive dreier Simulationsszenarien und zweier Softwarepakete zur hierarchischen Modellierung. Die Simulationsergebnisse zeigen, dass drei Gewichtungsansätze am besten geeignet sind, um die wahren Populationsparameter zu schätzen. Einer von ihnen beinhaltet nur die Verwendung von Gewichten der Ebene zwei (hier: Gewichte der Schulebene) und ist aufgrund seiner einfachen Umsetzung die Variante, die zu präferieren ist.

Acknowledgments

First, I would like to thank Prof. Dr. Helmut Küchenhoff for the opportunity to write my dissertation under his supervision. For almost 20 years now, I have greatly appreciated the honest, supportive, efficient, and constructive exchanges throughout my academic years, my professional work, and now in my dissertation. I sincerely look forward to and would be very grateful if the professional exchange would continue even after the doctoral thesis.

I want to thank Dr. Sabine Meinck no less warmly for her tireless and sympathetic support for co-authoring both papers and, not least, supporting my doctoral thesis. I am very grateful for your always open ear and for the constant exchange not to have lost the view of the economic aspects in the tunnel of scientific work. I look forward to having many more collaborations and cooperations in the future.

I also cordially thank Prof. Dr. Manfred Prenzel for his support and motivating words for the first paper. I admire your knowledge, empathy, and purposeful words at all times. Not least because of our same date for birthday, I hope the enriching exchange will continue in the future.

I would also like to thank Prof. Dr. Leslie Rutkowski very warmly for agreeing to examine my thesis. I am pleased about this and really appreciate it.

Lastly, I would like to thank my family for their support in completing my thesis. Thank you, Dominik, for always supporting me in doing this work and taking over our two little daughters without discussion when I needed time for the dissertation, even though you are very involved in your job yourself. Thank you also to my parents. You are always there for me and my family without hesitation supporting us in every situation. Finally, I like to thank my two daughters. Although for you it was a bit difficult to understand what your mother writes here, you mostly showed understanding of it.

Contents

Chapter 1: Overview

Introduction

Since its first assessment in the year 2000, PISA has established itself as a well-respected educational monitoring study in more than 80 participating countries worldwide. PISA measures students' basic competencies at the end of their secondary education. Empirical data are available in the key areas of *reading*, *mathematics,* and *science*, allowing conclusions about the performance of the participating states' education systems. In addition, the long-term established cross-sectional PISA studies at three-year intervals make it possible to identify and describe developments and transitions over time (Reiss et al., 2019).

Drawing a sample for PISA is demanding (OECD, 2017). The PISA international sampling design uses features attributed to "complex" samples. In the first selection stage, schools are sampled according to the *Probability Proportional to Size (*PPS; Meinck, 2020; Skinner, 2014) method, which implies larger schools have a higher probability of being sampled than smaller ones. Furthermore, stratification is applied. Explicit stratification means dividing all eligible schools (those with fifteen-year-old-students) into subgroups, with all schools belonging to a subgroup treated as a single sampling frame. Implicit stratification means sorting those separate frames by specific characteristics (Meinck, 2020). It differs from simple random sampling (SRS) as systematic sampling is applied to those ordered frames. The precision of the resulting estimates is similar to the results from proportional allocation and therefore this procedure is called implicit stratification in contrast to explicit stratification (Aßmann et al., 2011). The selected characteristics for stratification should be chosen to increase the estimator's efficiency compared to simple random sampling (Jaeger, 1984). In addition, international project management requirements and relevant privacy areas must be considered. Finally, the number of strata is also methodologically limited by sample size and the *Balanced Repeated Replication* (BRR) method (Valliant et al., 2018a).The research project *Evaluating German PISA stratification designs: a simulation study* aims to provide evidence for possible

improvements of the PISA stratification scheme of the German PISA sample. It may serve as a template for similar studies in other countries and economies participating in LSAs.

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 across several schools 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 considered in such models, which otherwise have to be reflected using special complex estimation procedures (e.g., BRR in PISA; OECD, 2017). Although there is sufficient evidence that sampling weights must be used in multilevel modelling (MLM) to obtain unbiased estimates (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. The main goal of the research project, *Sampling weights in multilevel modelling: an investigation using PISA sampling structures,* is to provide a clear recommendation for using weights and estimation procedures for multilevel analyses in LSAs.

PISA Sampling Design

Sampling procedures that allow for undistorted and precise population estimates must be applied to enable conclusions from the sample-based PISA assessment on the population of fifteen-year-old-students and ensure international comparability. In PISA, a state-of-the-art sampling design acknowledged by the scientific community is utilized (Rutkowski et al., 2013). PISA implements, by default, a complex sample design with a two-stage sampling procedure. As a rule, schools are drawn in the first stage, and students in the participating schools are systematically randomly selected in the second stage. A set of tightly scrutinized sampling standards assures this design in all participating countries (OECD, 2020).

PISA's internationally specified target population consists of all students in an age cohort. This is all fifteen-year-old students who attend the seventh or a higher grade. The exact definition of the age cohort is determined in coordination with the international PISA consortium and may vary slightly

between countries and economies due to different survey periods. For example, in Germany, all students born between January 1, 2002 and December 31, 2002 (inclusive) and attending at least grade 7 or higher were eligible to participate in PISA 2018. To implement the first sampling stage, a so-called school sampling frame is created – a comprehensive list of all schools where fifteen-year-old students are expected to be taught during the data collection period.

Within this list, school-level stratification is implemented. One can draw independent probability samples from each stratum by dividing the population into H non-overlapping subpopulations, called strata (Groves, 2011). Accordingly, a stratified random sample comprises of several subsamples, each representing internally more homogeneous subpopulations concerning the stratification characteristics. To make conclusions about the full population, the individual sample values must be weighted according to the ratios of the strata to the population. In stratified sampling, what matters is the variation within the strata. The strata should be determined such that the variables of interest within a stratum are as invariant as possible. In contrast, the different strata should differ as much as possible from each other to improve sampling efficiency (Jaeger, 1984; Lohr, 1999) and sampling precision (Cochran, 1977). In other words, it increases the sampling precision and results in smaller sampling errors of these variables (Cochran, 1977; Meinck & Vandenplas, 2021). Stratification information must be available for all eligible schools in the sampling frame. Using this information, the sampling frame can be sorted by the stratification variables before sampling. Requirements at the international level and national political sensitivity (such as the request for a fair regional distribution of the sample) may also play a role in the stratification. The variance between strata does not contribute to the variance of the estimator. Only the sample size proportional to its stratum size ensures that the sample will highlight the differences between strata. Estimating the sampling variance for stratified samples with SRS within the strata is straightforward and can be handled, e.g., via a variance decomposition. For complex samples such as those applied in PISA, estimation of sampling variance becomes more complicated as clustering effects and varying selection probabilities have to be accounted for within each stratum. Stratification can be implemented in two different

ways: explicit and implicit (OECD, 2020). Explicit stratification means a mutual grouping of the schools by specific school characteristics and sampling schools for each explicit stratum separately (Singh & Mangat, 1996). In the literature, this explicit stratification refers to the upper definition of stratified sampling (Lohr, 1999; Singh & Mangat, 1996; Thompson, 2012). Implicit stratification can be added within explicit strata, implying the sorting of the schools by further characteristics. The goal of this sorting is to approximately preserve the population proportions in the sample.

Furthermore, the PPS sampling procedure is applied (Meinck, 2020; Skinner et al., 1989) by having larger schools sampled with higher probability than smaller ones, and vice versa (Lohr, 1999). This procedure was first advocated by Mahalanobis (1952) and subsequently discussed by many researchers, e.g., Hansen and Hurwitz (1943) or Sukhatme et al. (1984). If the school size is used as the *Measure of Size* (MOS), i.e., the estimated number of fifteen-year-old students in a school, in PPS, larger schools have a higher probability of being sampled than smaller ones, and vice versa, as students within larger schools have smaller selection probabilities than students within smaller schools (Lohr, 1999). Selecting schools with varying probabilities will result in unbiased estimators if they are appropriately weighted according to their selection probabilities (Singh & Mangat, 1996). The size variable must be available in the sampling frame. In PISA, the preferred MOS is the expected number of fifteen-year-old students in each school. Other size measures, such as the total school size or the number of students in the modal grade, could be used as alternatives (OECD, 2020). The selection probability for a school i can then be written as

$$\pi_i = n\frac{MOS_i}{\sum_{j=1}^{N} MOS_j}, \tag{1}$$

with $nMOS_i < \sum_{j=1}^{N} MOS_j$, i being the selected schools, N being all schools in the population, and n being the sample size. For the variance of any estimator, the variation of the values in the sum is decisive (Lohr, 1999). This also shows the advantage of PPS sampling: if the variance of the calculated statistic in a school is higher than its division by the MOS of the respective school, the estimator has a

smaller sampling variance. This is met if MOS is proportional to the used statistic (Kauermann & Küchenhoff, 2011).

Sampling weights are provided as inverse selection probability to avoid bias due to disproportional selection probabilities (OECD, 2017). Those weights are computed as the inverse of the selection probabilities of each selection stage, adjusted for nonresponse. To account for the complex sampling design, standard errors must be estimated by respective statistical methods (Lohr, 1999). For computing unbiased estimates of the sampling variance, BRR (Wolter, 2007) with Fay's adjustment is used in PISA (Judkins, 1990). The advantage of BRR, but also similar replication methods like the Jackknife Repeated Replication (JRR), is that it can account for the effects on variances of nonresponse adjustments (as long as weighting steps are computed separately for each replication; Valliant et al., 2018b). However, this method is preferred over other methods, such as JRR, as it provides more stable estimates when analysing sparse population subgroups (Judkins, 1990; OECD, 2017; Rao & Shao, 1999). Specifically, if the estimate is a ratio of two subgroups, some replicate ratio estimates can be extremely large or undefined because of near-zero or undefined denominators, respectively (Rao & Shao, 1999; Rao & Wu, 1985).

Hierarchical Models

For LSA studies, it can be essential to consider student characteristics in the context of groups to which students belong. The clustered nature of these samples means that students within the same classes and schools are less likely to be independent of each other, as their knowledge, skills, and other attributes may be influenced by factors such as their classmates, teachers, school principals, and the overall school environment (Karakolidis et al., 2022; Raudenbush & Bryk, 2002). In this case, one characteristic affects individuals in two dimensions (or two levels of a hierarchy), i.e., at the individual and group levels. In addition, interactions between the two (or even more) levels are also possible and need to be considered (Meinck & Vandenplas, 2012).

To analyse LSA data, simple linear regressions are often used. However, these have the weakness of assuming that students' responses are independent of their group or school (Burstein, 1980). The

second difficulty is that often the assumption is made that the correlations within groups are the same as between those groups (i.e., most schools in LSAs). Those misalignments can lead to aggregation bias, fallacy (Cronbach, 1976), and underestimated precision (Aitkin et al., 1981; Meinck & Vandenplas, 2012; Woltman et al., 2012). To account for the hierarchical structure of LSA data, hierarchical linear models (HLM)[1] have been developed (Aitkin & Longford, 1986; De Leeuw & Kreft, 1986; Snijders & Bosker, 2012). Those models have the property to measure effects, relationships, and variability at different levels (Meinck & Vandenplas, 2012). Menezes et al. (2016) contend that - compared to simple linear modelling - hierarchical analysis allows for a more nuanced dissection of education assessment data while accounting for the many possible impact levels relevant to meaningful education policy. Because clustering students in classes and/or schools is a given fact in schools, it should be considered in both relevant analyses and policy decision-making processes (Karakolidis et al., 2022).

To enable statistical inference using hierarchical models (i.e., inferring from a sample on an infinite population), Pfeffermann et al. (1998) and Asparouhov (2006) argued that it is essential to include complex sampling designs, like those applied in PISA, in the hierarchical models. The so-called pseudo maximum likelihood (PML) estimation technique was developed by Skinner (1989), following the idea of Binder (1983), and includes sampling weights into the HLM analysis procedure. It defines a hybrid approach combining design-based and model-based inference estimation techniques.

The PML estimation technique applies the principles of the Horwitz-Thompson (HT) theorem by using the inverse of the selection probabilities as weights (Horvitz & Thompson, 1952)

$$\hat{Y}_{HT} = \frac{1}{N}\sum_{j=1}^{n} w_j\, y_j = \frac{1}{N}\sum_{j=1}^{n} \frac{1}{\pi_j} y_j, \qquad (2)$$

[1] Alternative naming of HLM are multilevel model, variance component model and random coefficient model.

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 $w_{ij} = \frac{1}{\pi_{ij}}$. To achieve a sum instead of the product for easier mathematical handling, the census log-likelihood follows

$$l\,(Y|\theta) = \sum_{i=1}^{m} w_i \log \int_{bi} exp\left[\sum_{j=1}^{n_i} \log w_{ij} f\,(Y_{ij}|\theta)\right] \phi\,(b_i)\partial b_i \tag{3}$$

with weights $w_i = \frac{1}{\pi_i}$ as inclusion probability for school i and $w_{ij} = \frac{1}{\pi_{ij}}$ as inclusion probability for student j given school i. The HT theory with 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). Setting the first derivation of this pseudo log-likelihood function to zero achieves the pseudo maximum likelihood estimator

$$\sum_{i=1}^{m} w_i \frac{\int_{bi}\left[exp\ \sum_{j=1}^{n_i} w_{ij} \log f\,(Y_{ij}|\theta)\right] \cdot \left[\sum_{j=1}^{n_i} \frac{\partial\ log f\left(Y_{ij}|\theta\right)}{\partial\theta}\right] \phi\,(b_i)\partial b_i}{\int_{bi}\left[\exp \sum_{j=1}^{n_i} w_{ij}\ \log f\,(Y_{ij}|\theta)\right]\ \phi\,(b_i)\partial b_i} \tag{4}$$

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).

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. The research project in this thesis, *Sampling weights in multilevel*

modelling: an investigation using PISA sampling structures, provides a clear recommendation for using weights and estimation procedures for multilevel analyses in LSAs. Besides the estimate, 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 and often too small. One way to deal with this is to use sandwich standard errors, a function of the modelled standard errors and observed residuals.

Simulation Approach

For the research projects *Evaluating German PISA stratification designs: a simulation study* and *Sampling weights in multilevel modelling: an investigation using PISA sampling structures,* it is necessary to reproduce the sampling frame and the associated target population as accurately as possible with all their characteristics. Samples are then repeatedly drawn from these populations using a Monte Carlo simulation (Boulesteix et al., 2020; Morris et al., 2019; Thomopoulos, 2013), and accordingly, research questions of these projects are answered and discussed under theoretical consideration with findings from these simulations studies. Regarding the literature, two widely used simulating population methods will be described in the following.

In the first approach, weights of an existing sample can be disposed such that this simulation approximates the actual population. The basis of this approach has been developed by Little (1993) and Rubin (1993), discussed by Beckman et al. (1996) and developed in recent applications like in Templ et al. (2017). We use the student sample of the German PISA 2018 data as a basis for the simulation (Reiss et al., 2021). By aggregating student data (using school identifiers) to the level of schools, we achieve a school dataset. As the true anonymous list of schools from PISA 2018 with information on the number of PISA eligible students is available to the authors, we add information on the school's MOS, federal state, and school type to the data. We did not only use information from the list of schools because other characteristics, such as student achievement and migration

background, are available in the sample. To simulate the German school frame using a sample of schools, each school has been copied according to its (rounded) school weight. A school from the sample then represents several schools according to their weight in the population. For example, a sampled school with a school weight of 10.21 was copied 10 times on the simulated school frame as it represents about 10 other schools in the population. This approach gives us an approximated copy of the complete school frame and is applied in the project *Evaluating German PISA stratification designs: a simulation study.*

The second simulation approach is generated using the properties of the desired characteristics and their correlation among an existing distribution assumption (Mang et al., 2021). Thus, it is based on two data sources. The first source is the sampling frame of a specific assessment. Further, relevant population features were estimated based on the sample of the same assessment as the abovementioned sampling. Those characteristics are then added to each school of this 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, 1995) 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. This simulating population method is discussed in the project *Sampling weights in multilevel modelling: an investigation using PISA sampling structures*.

Chapter 2: Summary and Conclusions

Since the beginning of LSA studies dates back to the 1960s (Husén & Postlethwaite, 1996), when "measured outcomes and their determinants within and between systems of education" were first discussed (Karakolidis et al., 2022), many methodologies for sampling and evaluating LSA studies have been developed and established. Several of these methods were developed when it was impossible to check their evidence and determine their quality by simulation (Boulesteix et al., 2020) and thus evaluate them if necessary. In addition, there are constant theoretical new methodological developments and adjustments that have to be taken into account on a continuous progression. On that basis clear implications or improvements for further studies and analyses can be made (Morris et al., 2019).

Regarding a possible optimization of the stratification design in PISA, the study shows that a change in stratification can be suggested for some of the presented approaches. However, it turned out that the authors' new derived approach having school types as explicit and federal states as implicitly given stratification (with special handling of the federal state of Saarland) is highly favorable as it considers all relevant aspects of a possible change. Those are particularly the available information in official statistical offices. To be able to construct the sampling frame, the improvement of this approach is decreasing the standard error of any statistics given by evidence through the simulation study and the conservative and cautious change by preserving the structure of Germany with its federal states and school types. Thus, it follows a reasonable communication of this change to relevant policymakers and stakeholders (for example, the press or teacher unions).

The study for weighting in hierarchical models for LSA data has shown that the authors' new derived approach using only the school weights provides the most unbiased estimates for hierarchical models. In this scenario, the final school weights are specified at the school level, while no weight is used at the student level. Final school weights reflect the school selection probabilities, adjusted for school nonresponse, and are typically provided with the public datasets of LSA. This recommended

weighting approach will help many researchers apply MLM with weights, thus driving further insightful research in the field of LSA.

Finally, it should be emphasized that the effects of optimizing stratification in the sampling of PISA will also be visible in the magnitude of the standard error when analysing large-scale assessment data in hierarchical modelling using sampling weights. However, this outcome does not change the suggestion of using only school level weights for the analyses. Overall, it can be pointed out that the presented studies gain insights into efficiently improving methods regarding PISA sampling structures and their application in weighting.

References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical Modelling of Data on Teaching Styles. *Journal of the Royal Statistical Society. Series a*(144 (4)), 419–461.

Aitkin, M., & Longford, N. (1986). Statistical Modelling Issues in School Effectiveness Studies. *Statistical Modelling Issues in School Effectiveness Studies*(149 (1)), 1-43.

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

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). *4 Sampling designs of the National Educational Panel Study: challenges and solutions* (Vol. 14). https://doi.org/10.1007/s11618-011-0181-8

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part a: Policy and Practice*, *30*(6), 415–429. https://doi.org/10.1016/0965-8564(96)00004-3

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

Boulesteix, A.-L., Groenwold, R. H., Abrahamowicz, M., Binder, H., Briel, M., Hornung, R., Morris, T. P., Rahnenführer, J., & Sauerbrei, W. (2020). Introduction to statistical simulations in health research. *BMJ Open*, *10*(12), e039921. https://doi.org/10.1136/bmjopen-2020-039921

Burstein, L. (1980). The Analysis of Multilevel Data in Educational Research and Evaluation. *The Analysis of Multilevel Data in Educational Research and Evaluation*(8), 158–233.

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

Cochran, W. G. (1977). *Sampling techniques* (3. ed.). *A Wiley publication in applied statistics*. Wiley.

Cronbach, L. J. (1976). How can instruction be adapted to individual differences? In R. M. Gagné (Ed.), *Learning and individual differences.* Merrill Books.

De Leeuw, J., & Kreft, I. (1986). Random Coefficient Models for Multilevel Analysis. *Journal of Educational and Behavioral Statistics*, *11 (1)*, 57–85.

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

Groves, R. M. (2011). *Survey Methodology* (2nd ed (Online-Ausg.)). *EBL-Schweitzer: v.561*. Wiley.

Hansen, M. H., & Hurwitz, W. N. (1943). On the Theory of Sampling from Finite Populations. *The Annals of Mathematical Statistics*, *14*(4), 333–362. https://doi.org/10.1214/aoms/1177731356

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

Husén, T., & Postlethwaite, T. N. (1996). A Brief History of the International Association for the Evaluation of Educational Achievement (TEA). *Assessment in Education: Principles, Policy & Practice*, *3*(2), 129–141. https://doi.org/10.1080/0969594960030202

Jaeger, R. M. (1984). *Sampling in education and the social sciences*. Longman.

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*(No. 6), 223–239.

Karakolidis, A., Pitsia, V., & Cosgrove, J. (2022). Multilevel Modelling of International Large-Scale Assessment Data. In M. S. Khine (Ed.), *Methodology for Multilevel Modeling in Educational Research : Concepts and Applications* (pp. 141–159). Springer Singapore Pte. Limited. https://doi.org/10.1007/978-981-16-9142-3_8

Kauermann, G., & Küchenhoff, H. (2011). *Stichproben*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12318-4

Kish, L. (1995). *Survey Sampling.* John Wiley and Sons.

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

Little, R. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, *Vol. 9 No. 2*, 407–426.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, *9*(8). https://doi.org/10.18637/jss.v009.i08

Mahalanobis, P. C. (1952). Some aspects of the design of sample surveys. *The Indian Journal of Statistics*(12), 1–7.

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education*, *9*(1), 1–39. https://doi.org/10.1186/s40536-021-00099-0

Meinck, S. (2020). Sampling, Weighting, and Variance Estimation. In H. Wagemaker (Ed.), *IEA Research for Education, A Series of In-depth Analyses Based on Data of the International Association for the Evaluation of Educational Achievement (IEA). Reliability and Validity of*

International Large-Scale Assessment: Understanding IEA's Comparative Studies of Student Achievement (1st ed., pp. 113–129). Springer International Publishing. 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.*

Meinck, S., & Vandenplas, C. (2021). Sampling Design in ILSA. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings* (pp. 1–25). Springer International Publishing. https://doi.org/10.1007/978-3-030-38298-8_25-1

Menezes, I. G., Duran, V. R., Mendonça Filho, E. J., Veloso, T. J., Sarmento, S. M. S., Paget, C. L., & Ruggeri, K. (2016). Policy Implications of Achievement Testing Using Multilevel Models: The Case of Brazilian Elementary Schools. *Frontiers in Psychology*, *7*, 1727. https://doi.org/10.3389/fpsyg.2016.01727

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Muthén, L. K., & Muthén, B. O. (2017). Mplus User´s Guide: Eighth Edition (Los Angeles, CA: Muthén & Muthén).

OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing.

OECD. (2020). *PISA 2018 Technical Report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

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

R Core Team. (2018). *R: A language and environment for statistical computing [*Computer software]. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

R Core Team. (2020). *R: A language and environment for statistical computing [*Computer software]. R Foundation for Statistical Computing. Vienna, Austria.

Rao, J. N. K., & Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, *86*(2), 403–415. https://doi.org/10.1093/biomet/86.2.403

Rao, J. N. K., & Wu, C. F. J. (1985). Inference From Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics. *Journal of the American Statistical Association*, *80*(391), 620. https://doi.org/10.2307/2288478

Raudenbush, S., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). *Advanced quantitative techniques in the social sciences series: Vol. 1*. Sage Publications.

Reiss, K., Mang, J., Heine, J.-H., Weis, M., Schiepe-Tiska, A., Diedrich, J., Klieme, E., & Köller O. (2021). *Programme for International Student Assessment 2018 (PISA 2018). Dataset*. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_PISA_2018_v1

Reiss, K., Weis, M., Klieme, E., & Köller, O. (Eds.). (2019). *PISA 2018: Grundbildung im internationalen Vergleich. PISA 2018*. Waxmann. https://doi.org/10.31244/9783830991007

Rubin, D. B. (1993). Discussion statistical Disclosure Limitation. *Journal of Official Statistics*(Vol.9 No. 2), 461–468.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of International Large-Scale Assessment*. Chapmall Hall/CRC; Chapman and Hall/CRC. https://doi.org/10.1201/b16061

SAS Institute Inc. (2018). *SAS-STAT Software* (Version 9.4) [Computer software]. Cary, NC. http://www.sas.com/

Singh, R., & Mangat, N. S. (1996). *Elements of Survey Sampling* (Vol. 15). Springer-Science+Business Media, B.V.; Springer Netherlands. https://doi.org/10.1007/978-94-017-1404-4

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.

Skinner, C. J. (2014). Probability Proportional to Size (PPS) Sampling. *Wiley StatsRef: Statistical Reference Online*, 1–5. https://doi.org/10.1002/9781118445112.stat03346.pub2

Skinner, C. J., Holt, D., & Smith, T. (Eds.). (1989). *Wiley series in probability and mathematical statistics : Applied probability and statistics*. *Analysis of complex surveys*. Wiley.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE Publishing.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., & Asok, C. (Eds.). (1984). *Sampling Theory of Surveys with Applications*. Iowa State University Press; Ames and Indian Society of Agricultural Statistics.

Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of Synthetic Complex Data: The R Package simPop. *Journal of Statistical Software*, *79*(10). https://doi.org/10.18637/jss.v079.i10

Thomopoulos, N. T. (2013). *Essentials of Monte Carlo simulation: Statistical methods for building simulation models*. Springer. https://doi.org/10.1007/978-1-4614-6022-0

Thompson, S. K. (2012). *Sampling* (3. ed.). *Wiley series in probability and statistics*. Wiley. https://doi.org/10.1002/9781118162934

Valliant, R., Dever, J. A., & Kreuter, F. (Eds.). (2018a). *Practical Tools for Designing and Weighting Survey Samples*. Springer, Cham.

Valliant, R., Dever, J. A., & Kreuter, F. (2018b). Variance Estimation. In R. Valliant, J. A. Dever, & F. Kreuter (Eds.), *Practical Tools for Designing and Weighting Survey Samples* (pp. 421–480). Springer, Cham. https://doi.org/10.1007/978-3-319-93632-1_15

Wolter, K. M. (2007). *Introduction to Variance Estimation* (2nd ed.). *Springer series in statistics*. Springer New York.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52–69. https://doi.org/10.20982/tqmp.08.1.p052

Chapter 3: Evaluating German PISA stratification designs: a simulation study

Chapter 3 determines which stratification scheme is best for PISA in Germany. For this, we examine the complex sampling design on theoretical considerations and discuss and examine these proposals in a simulation-based setting to substantially improve the sampling of the PISA study. Furthermore, findings about the estimation accuracy of different standard errors and constraints imposed by the international sampling design, the available information about schools, and specific national characteristics of the German educational system were evaluated. We verify seven different stratification designs. The selection is based on scenarios used in past LSAs in Germany and theoretical, new devised considerations for future implementations. The chosen scenarios were compared with two reference scenarios, (1) an unstratified design and (2) a synthetic optimal stratification design. The software program R 4.1.0 (R Core Team, 2020) were used for simulating the sample replicates. The analyses to quantify the differences between those stratification methods were also performed with R and the package survey (Lumley, 2004).

Contributing article:

Mang, J., Küchenhoff, H. & Meinck, S. (2023). *Evaluating German PISA stratification designs: a simulation study.* Revised manuscript submitted for publication.

Author contributions:

Mang devised the theoretical considerations, the research questions, derived and implemented the simulation study, conducted the data analysis, and wrote the first draft of the manuscript. All authors contributed to the interpretation of the results and the writing and revision of the manuscript.

Evaluating German PISA stratification designs: a simulation study

Julia Mang[1]*, Helmut Küchenhoff[2], Sabine Meinck[3]

[1] Technical University of Munich (TUM), TUM School of Social Sciences and Technology, Centre for International Student Assessment (ZIB), Munich, Germany

[2]Ludwig-Maximilians-Universität München, Institut für Statistik, Munich, Germany

[3]International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany

***Correspondence:**

Julia Mang

Technical University of Munich (TUM), TUM School of Social Sciences and Technology, Centre for International Student Assessment (ZIB)

Arcisstr. 21

Munich, 80333

Germany

Julia.Mang@tum.de

Abstract

Stratification is an important design feature of many studies using complex sampling designs and it is often used in large-scale assessment (LSA) studies, such as the *Programme for International Student Assessment* (PISA), for two main reasons. First, stratification variables that achieve a high between and low within strata variance can improve the efficiency of a survey design. Second, stratification allows one to, explicitly or implicitly, control for sample sizes across subpopulations. It ensures that some parts of a population are in the sample in predetermined proportions.

In this study, we determine through simulation which stratification scheme is best for PISA in Germany. For this, we consider the constraints imposed by the international sampling design, the available information about schools, and specific national characteristics of the German educational system. We examine seven different stratification designs selected based on scenarios used in past LSAs in Germany and theoretical considerations for future implementations. The chosen scenarios were compared with two reference scenarios: (1) an unstratified design and (2) a synthetic optimal stratification design.

The simulation study reveals that the stratification design currently applied in PISA produces satisfactory results regarding sampling precision. The present stratification design is based on Germany's federal states and school types. However, this approach leads to small strata, which has been problematic for estimating sampling variance in previous cycles. Therefore, alternative stratification scenarios were considered and, in addition to overcoming the small-strata problem, also led to smaller standard errors for estimates of student mean performance in mathematics, science, and reading.

As a result of this study, we recommend considering three different stratification designs for Germany in future cycles of PISA. These recommendations aim to: (1) improve the sampling efficiency while keeping the sample size constant, (2) follow a sound methodological approach, and (3) make conservative and cautious changes while maintaining a reflection of the structure of the German federal school system with different school types. These suggestions include a reinvented stratification of grouped German federal states and designs with school types as explicit stratifiers and federal states as implicit stratifiers.

Keywords

Programme for International Student Assessment (PISA), large-scale assessment (LSA), stratification, explicit stratification, implicit stratification, systematic random sampling, simulation study, sampling weights

Introduction

Drawing a sample for the *Programme for International Student Assessment* (PISA) to represent the target population of fifteen-year-old students is demanding (OECD, 2017). The PISA international sampling design uses features attributed to "complex" samples. The overall design can be described as a stratified two-stage random sample. In the first selection stage, schools are sampled with a *probability proportional to their size* (PPS; Meinck, 2020; Skinner, 2014), which implies that larger schools have a higher probability of being sampled relative to smaller schools. In a second stage, about 30 to 40 fifteen-year-old students are systematically randomly sampled across participating schools with equal probabilities after sorting them by gender and grade. Such a selection procedure is also called cluster sampling.

School-level stratification can be implemented in two different ways. Explicit stratification involves dividing all eligible schools (those with fifteen-year-old students) into subgroups, with all schools belonging to a subgroup treated as a single sampling frame. Implicit stratification means sorting those separate frames by specific characteristics (Meinck, 2020).It differs from simple random sampling (SRS) as systematic sampling is applied to those ordered frames. The precision of the resulting estimates is similar to the results from proportional allocation and therefore this procedure is called implicit stratification in contrast to explicit stratification (Aßmann et al., 2011). Stratification improves the efficiency of the sampling design if the variables used for stratification are correlated with the variables of interest (e.g., mean student proficiency). In other words, it increases the sampling precision and results in smaller sampling errors of estimates of these variables (Cochran, 1977; Meinck & Vandenplas, 2021) if the variance between the strata becomes large and the variance within the strata is small. It further ensures that some parts of the population are included in the sample in predetermined proportions. With implicit stratification, the proportions in the population are approximately preserved in the sample. Explicit stratification, however, allows for a disproportional sample allocation.

Sampling weights and nonresponse adjustments are provided to avoid bias due to disproportional selection probabilities that combine the inverse selection probabilities at each sampling stage with nonresponse adjustments (OECD, 2017). Using them with the Horvitz-Thompson (HV) estimator allows for unbiased and consistent estimators for any desired statistic. For computing unbiased estimates of the sampling variance accounting for the complex design, *Balanced Repeated Replication* (BRR) with Fay's adjustment is used (Judkins, 1990). To implement this method, pairs of primary sampling units (usually schools) are created based on their location in the sorted sampling frame within each explicit and implicit stratum, whenever possible (OECD, 2017). That is, schools in one pair, also called a "variance zone", are those sampled schools next to each other in the sampling frame, thereby sharing specific characteristics as they belong to the same stratum. Replicate weights are then calculated using a specific re-weighting scheme to accommodate the BRR computation algorithm (OECD, 2017; Rust & Rao, 1996).

Determining an efficient stratification scheme in international large-scale assessments in education (LSA) is not trivial. The selected characteristics for stratification should be chosen to increase the estimator's efficiency compared to simple random sampling (Jaeger, 1984). In addition, international project management requirements and relevant privacy areas must be considered. Finally, the number of strata is also methodologically limited by the sample size and the BRR method (Valliant et al., 2018a). This study aims to provide evidence aimed at supporting the improvement of the stratification design used for the German sample in PISA. It may serve as a template for similar studies in other countries and economies participating in LSA.

In previous PISA cycles, the German sample has been stratified using federal states as an explicit stratification variable with 16 categories and school type as an implicit stratification variable (Mang et al., 2019). When preparing school nonresponse adjustments for this sampling scheme in previous rounds of PISA, it was found that some strata could become very small or even empty. During the school nonresponse adjustment, initial adjustment cells are based on explicit and implicit stratification variables. School-level nonresponse or school closures could induce very small

adjustment cells. For example, in 10 out of 16 federal states (62.5%), fewer than 10 schools were selected in PISA 2018. Because small cells can lead to unstable weight adjustments and, in turn, inflate the sampling variances, it is a common practice to collapse small adjustment cells. These collapsed strata no longer accurately reflect the implemented sampling design, likely inflate the within strata variance, and show smaller efficiency gains compared to simple random samples when computing standard errors (SEs). Furthermore, federal states may not be effective predictors of achievement since many states share similar average achievement levels and variances within those strata might be too large to result in smaller sampling variances. Thus, other variables like the proportion of students with migration backgrounds within schools or students' average socioeconomic background may be more closely related to achievement and, therefore, could be preferred stratification variables (Buchmann & Park, 2009).

This study examines how different stratification designs of the German PISA sample can lead to an increase in precision in estimating the main outcome variables: student performance in mathematics, science, and reading. We aim to identify and recommend a stratification design that aligns with both international and national requirements, is feasible in terms of its practical implementation, and is highly efficient. Since the results of the PISA study enjoy great publicity in Germany and are closely examined by politicians and the press, it is important to both use an unbiased and efficient estimation as well as be able to communicate design changes to a non-technical audience effectively. We focus on three schemes that will be benchmarked against a design without stratification and an artificial "perfect" stratification. Comparisons of the current design and the proposed alternatives will be made to quantify the differences between them and thus, support recommendations for a change in stratification with evidence.

This paper is organized as follows. The first section elaborates on the PISA sampling design with PPS sampling, the stratification process, and its application in the German sample. Next, we introduce the simulation study. This section describes the process of simulating the PISA population and the process of stratification, sampling, and creating estimation weights for the analyses. We then

describe the performed analyses to compare and quantify the different stratification designs. Afterwards, we present and discuss the simulation study results, determining the differences and benefits that can result from different stratification designs and providing our recommendations for future data collections. Finally, we discuss the generalizability of our findings and possibilities for future research.

Design-based multistage sampling in PISA

PISA collects data from a multistage sample of fifteen-year-old students in all participating countries and economies. For this purpose, probabilistic random samples are selected, which can be used to generalize on the population, for example, to all schools having fifteen-year-old students in Germany (Brown, 2010; Kish, 1965; Levy & Lemeshow, 2013; Thompson, 2012). To make correct inferences about the population of fifteen-year-old students in school and to ensure international comparability, sampling procedures in PISA must be applied that allow for undistorted and precise population estimates. Special attention is paid to the point estimate of the characteristic of interest and its precision (Meinck, 2020). In PISA, a state-of-the-art sampling design acknowledged by the scientific community is applied (Rutkowski et al., 2013). PISA implements, by default, a complex sample design with a two-stage sampling procedure. As a rule, schools are drawn in a first stage, and students in participating schools are systematically randomly selected in a second stage.

PISA's internationally specified target population consists of all students in an age cohort. This is, generally, all fifteen-year-old students who attend grade 7 or higher. The exact definition of the age cohort is determined in coordination with the international PISA consortium and may vary slightly between countries and economies due to different survey periods. For example, in Germany, all students born between January 1, 2002 and December 31, 2002 (inclusive) and attending at least grade 7 or higher were eligible to participate in PISA 2018. A so-called school sampling frame is created to implement the first sampling stage. This is a comprehensive list of all schools where fifteen-year-old students are expected to be taught during the data collection period. The purpose of this frame is to provide a comprehensive list of all eligible primary sampling units (here: schools)

containing all units of the target population (here: 15-year-old students; Meinck, 2020).

In Germany, the information for this list is collected from the statistical agencies of the federal states. It includes, among other variables, the school type, the funding body, the number of students from the target population (7^{th} to 10^{th} grade, born in the year of definition), the number of 7^{th} to 10^{th} grade classes as well as information about planned school mergers or school closures. It should be emphasized that the information made available is mostly data protection insensitive according to GDPR[1], which is an important consideration when deciding how to design the sample.

In the PISA sampling frame, a school is defined as an organizational unit with one or more buildings belonging to that school. However, if a school has different tracks within that organizational unit, each track is listed separately. Within comprehensive schools or schools with several educational programs, the school track defines the intended school qualification of students in the associated branch. The German federal states partially define different tracks which can be divided into three different branches: lower secondary with no access to upper secondary (basic general education), lower secondary with access to upper secondary (extensive general education), and higher secondary (academic education). This definition forms the basis of school types for the stratification.

PPS Sampling

In PISA, the PPS sampling procedure is applied for the school selection (Meinck, 2020; Skinner, 2014). This procedure was first advocated by Mahalanobis (1952) and subsequently discussed by many researchers, e.g., Hansen and Hurwitz (1943) or Sukhatme et al. (1984). If the school size is used as the measure of size (MOS) in PPS, larger schools have a higher probability of being sampled than smaller ones, and vice versa, as students within larger schools have smaller selection probabilities than students within smaller schools (Lohr, 1999). Selecting schools with varying probabilities will result in unbiased estimators if they are appropriately weighted according to their selection probabilities (Singh & Mangat, 1996). The size variable must be available in the sampling frame. In

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1

PISA, the preferred MOS is the expected number of fifteen-year-old students in each school. Other size measures, such as the total school size or the number of students in the modal grade, could be used as alternatives (OECD, 2020). The selection probability for a school i can then be written as

$$\pi_i = n\frac{MOS_i}{\sum_{j=1}^{N} MOS_j}, \tag{1}$$

with $nMOS_i < \sum_{j=1}^{N} MOS_j$, i being the selected schools, N being all schools in the population, and n being the sample size. For the variance of any estimator, the variation of the values in the sum is decisive (Lohr, 1999). This also shows the advantage of PPS sampling: if the variance of the calculated statistic in a school is higher than its division by the MOS of the respective school, the estimator has a smaller sampling variance. This is met if MOS is proportional to the used statistic (Kauermann & Küchenhoff, 2011).

Sampling weights are provided to avoid bias due to disproportional selection probabilities (OECD, 2017). Those weights are computed as the inverse of the selection probabilities of each selection stage. Not all sampled schools and students eventually participate in the assessment. In Germany, the PISA assessment is mandatory for public schools, so they cannot reject participation. However, private schools do sometimes refuse to participate. At the student level, students may not participate in the test if they are sick on the assessment day or if they changed schools between the time of listing and assessment. In the event of such nonresponse, other "similar" students who participate (those belonging to the same gender and grade) carry the weight of their nonresponding peers. This avoids under-representation of those students. In short, nonresponse adjustment cells are built within each explicit stratum, grade, gender, and school combination (OECD, 2020). This nonresponse factor is thus, also considered in the sampling weights. The PPS method adjusts for nonresponse results by creating unequal weights in smaller sampling errors when estimating population features and increases the estimator's efficiency. Combined with systematic random sampling within schools, it is also called a self-weighting design (Solon et al., 2015). Moreover, PPS

sampling is a simple way to ensure similar final sampling weights when selecting an approximately equal number of students in each sampled school (Meinck, 2020).

Stratification

The word "stratify" comes from Latin word meaning “to make layers.” One can draw independent probability samples from each stratum by dividing the population into H non-overlapping subpopulations, called strata (Groves, 2011). Accordingly, a stratified random sample comprises of several subsamples, each representing internally more homogeneous subpopulations concerning the stratification characteristics. To make conclusions about the full population, the individual sample values must be weighted according to the ratios of the strata to the population. In stratified sampling, what matters is the variation within the strata. The strata should be determined such that the variables of interest within a stratum are as invariant as possible. In contrast, the different strata should differ as much as possible from each other to improve sampling efficiency (Jaeger, 1984; Lohr, 1999) and sampling precision (Cochran, 1977). Stratification information must be available for all eligible schools in the sampling frame. Using this information, the sampling frame can be sorted by the stratification variables before sampling. Requirements at the international level and national political sensitivity (such as the request for a fair regional distribution of the sample) may also play a role in the stratification. The variance between strata does not contribute to the variance of the estimator. Only the sample size proportional to its stratum size ensures that the sample will highlight the differences between strata. Estimating the sampling variance for stratified samples with SRS within the strata is straightforward and can be handled, e.g., via a variance decomposition. For complex samples such as those applied in PISA, estimation of sampling variance becomes more complicated as clustering effects and varying selection probabilities have to be accounted for within each stratum.

Stratification can be applied at any stage of the multistage sampling design. In PISA, two types of stratification are used: explicit and implicit (OECD, 2020). Explicit stratification means the grouping of schools by specific school characteristics and sampling schools for each explicit stratum separately

(Singh & Mangat, 1996). In the literature, explicit stratification is what is referred to in stratified sampling (Lohr, 1999; Singh & Mangat, 1996; Thompson, 2012). Implicit stratification can be added within explicit strata and involves the sorting of the schools by further characteristics. Combined with the PPS sampling approach methods, implicit stratification can be described as a systematic random PPS sampling design within each explicit stratum. The goal of this sorting is to approximately preserve the population proportions in the sample.

PISA establishes quality standards all participating countries and economies must adhere to. One of these standards specifies that schools must be sampled using agreed upon, established, and professionally recognized principles of probability sampling. One of these principles involves the identification of appropriate stratification variables to reduce sampling variance and facilitate the computation of nonresponse adjustments (OECD, 2020). Stratification schemes differ considerably across the participating educational systems. For instance, the OECD (2020, Table 4.1) lists the stratification schemes for all participating countries and economies in their technical report. Urbanization, ISCED levels, school funding, countries' and economies' languages, school types, school sizes, or school tracks have been chosen in the past as stratification categories. In addition, the percentage of school variance explained by explicit stratification variables by country and domain (OECD, 2020, Annex C1) differs widely between the participating countries and economies. The potential effects of an optimal stratification design can be illustrated using the example of the Netherlands in Tables C4 and C5 of the Annex of the OECD Technical Report for PISA 2018 (OECD, 2020). For example, the intraclass correlation (ICC) for the domain reading is 0.53. This means the variances between and within the schools are equally distributed between and within the schools. After considering stratification (explicit stratification in the Netherlands: school types, Table 4.1. of the Technical Report, OECD, 2020), it is only 0.10, i.e., the variance within schools barely plays a noteworthy role anymore. A similar effect of stratification on variance decomposition can also be observed for France.

To understand the current stratification design used for PISA in Germany, a look into the past may be helpful. In the first three cycles (2000, 2003, and 2006), PISA was used to facilitate comparisons between the German federal states, which comprise of independent educational school systems with independent governance. Explicit stratification and oversampling by federal states were necessary to accommodate this national requirement. Each federal state was treated as a separate population of interest. While not needed in later cycles, this stratification design was kept to simplify the communication of the results to the broader audience unfamiliar with the technicalities of complex samples. In addition, education policy representatives from the federal and state governments called for such a design, as it appropriately reflects and relates to the diversity of different education systems at the federal state level. In addition to explicit stratification by state, implicit stratification by school type has been implemented in each cycle. This ensures that sampled students were distributed as evenly as possible across Germany so that each combination of federal state and school type was represented with at least some minimum number of schools in the sample.

Germany applies different stratification designs in other LSAs, at least more recently. For example, in the *Progress in International Reading Literacy Study* (PIRLS) and the *Trends in International Mathematics and Science Study* (TIMSS), the German stratification design is based on an indicator of the socioeconomic background of students and school types (for more details, see Mullis et al., 2016, Chapter 5). The socioeconomic indicator has been determined by the number of students with an immigration background in each school eligible for the respective study.

Estimation procedures for multistage, stratified PPS sampling

To determine the correct estimation procedure for any survey statistic when complex sampling is applied, the characteristics of the sample design and the form of the required statistic must be considered (Wolter, 2007). The form of a statistic can be distinguished into linear and non-linear estimators. Those can be, for example, means from a straightforward sampling design or ratio estimators under complex sampling design. In detail, Wolter (2007) or Valliant et al. (2018a) provide

a theoretical background for those distinctions. The characteristics of the sampling design influence the precision measure of any statistic, in particular.

Horvitz-Thompson Estimator

The HV estimator can be used for any linear and non-linear statistic with the constraint that no element (i.e., the students in this context) can be sampled with replacement. The estimation formula can be written as

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{y_{ij}}{\pi_{ij}}, \tag{2}$$

with π_{ij} = the selection probability that the j-th student is selected within the i -th school, N being the number of students in the population, m and n being the number of schools and students in the sample, respectively. y_{ij} indicates the statistic from the students. The Horvitz-Thompson estimator weights the selected students within the schools chosen by their inverse selection probabilities π_{ij}. Thereby, the mechanism of the PPS sampling procedure is applied for the selection of the schools. This step is defined in this context as schools being the *Primary Sampling Units* (PSU). This estimator provides unbiased and consistent estimates for almost all linear and nonlinear statistics (Horvitz & Thompson, 1952), also known as the Horvitz-Thompson-theorem (Singh & Mangat, 1996).

Variance estimation

To account for the uncertainty in the estimation resulting from the complex sampling design, standard errors must be estimated by their respective statistical methods (Lohr, 1999). For computing unbiased and consistent estimates of sampling variance, the BRR method with Fay's adjustment is used in PISA (Judkins, 1990). The advantage of BRR, but also of similar replication methods like the Jackknife Repeated Replication (JRR), is that it can account for the effects on variances of nonresponse adjustments (as long as weighting steps are computed separately for each replication; Valliant et al., 2018b). However, this method is preferred over other methods, such as JRR, as it provides more stable estimates when analysing sparse population subgroups (Judkins, 1990; OECD, 2017; Rao & Shao, 1999). Specifically, if the estimate is a ratio of two subgroups, some

replicate ratio estimates can be extremely large or undefined because of near-zero or undefined denominators, respectively. (Rao & Shao, 1999; Rao & Wu, 1985). For proficiency estimates in PISA, standard errors are a combination of sampling and imputation errors. Still, this paper focuses only on the sampling error as the sampling error is generally much larger than the imputation error. Therefore, the imputation error can be neglected in the context of this paper (OECD, 2020).

To implement the method of BRR, pairs of primary sampling units (usually schools) are created according to the order of appearance in the sampling frame, which is first sorted by explicit strata, then by implicit strata and size (i.e., in Germany, first by the federal states, then by school types and size). Hence, schools in a pair often share similar characteristics, as they belong to the same stratum. Pairs are sequentially numbered and named as *variance zones (*or just simple *zones*); other common names are *variance strata* or *pseudo-strata*. One school within these pairs is randomly numbered as one, the other as two.

Then, 80 replicate weights are calculated using a specific re-weighting scheme to accommodate the BRR computation algorithm (OECD, 2017; Rust & Rao, 1996). That is, the estimation weight of each student within one school in the pair is multiplied by 1.5, while the estimation weight of each student in the other school in the pair is multiplied by 0.5. In cases where there are three units in a triplet, either one of the schools (designated at random) receives a factor of 1.7071 for a given replicate, with the other two schools receiving factors of 0.6464, or else the one school receives a factor of 0.2929 and the other two schools receive factors of 1.3536. Determining which schools receive inflated and deflated weights is carried out systematically, based on the entries in a Hadamard matrix of order 80 (OECD, 2017). This Hadamard matrix only contains the values -1 and 1, and multiplication with its transposed counterpart returns an identity matrix of order 80 multiplied by a factor 80 (Wolter, 2007). Technically, this is like selecting sub-samples from the whole sample, achieved by systematically manipulating the estimation weights. The PISA 2000 Technical Report (OECD, 2002, Appendix 12) explains how these particular factors came to be used. More than 80 replicates would

not improve the precision and would only add computational time. In addition, each replication weight is adjusted for nonresponse at both school and student levels.

Given the variance estimator for a specific analysed statistic named X^* from the full sample follows

$$\widehat{V_{BRR}}(X^*) = 0.05 \sum_{t=1}^{80} \{(X_t^* - X^*)^2\} \tag{3}$$

with t=1,…, 80 being the number of replicates. X_t^* results in the t[th] estimation of this statistic with the t[th] replication weights combination. The advantage of the BRR method is that it produces unbiased and consistent estimators under complex designs (OECD, 2017).

Research Questions

Utilizing a simulation study, we aim to answer the following research questions in this paper:

1) Are there relevant differences in the SEs and the bias of mean achievement estimates of specific PISA domains when applying different stratification schemes for school sampling?
2) What is the best stratification design for PISA Germany, considering suggestions from research question 1 and constraints determined by the international sampling design, the available information about schools, and specific national characteristics of the educational system?

Simulation Study

With the help of a simulation study, the most efficient stratification procedure that also complies with the abovementioned requirements should be identified. In detail, we compare schemes used in the past with schemes that show promise for providing more precise results, benchmarking them against both a scheme without stratification and a scheme reflecting a "perfect" stratification.

The simulated school population is based on the German PISA 2018 school population. From this "population", 2,000 sample replications are selected according to the stratification characteristics defined in the next section, using the approach of a Monte Carlo simulation. For each dataset, simulated weights and replication weights are calculated when drawing the sample for each stratification variant.

The software program R Studio Version 1.4.1717 (RStudio Team, 2020) and its corresponding program R 4.1.0 (R Core Team, 2020) were used for simulating the sample replicates. The analyses to quantify the differences between those stratification methods were also performed with R Studio, its corresponding program R and the package *survey* (Lumley, 2004).

Simulation PISA Population

The simulation of a population can be implemented using two different methods. First, it can be generated using the properties of the desired characteristics and their correlation with each other with an existing distribution assumption (Mang et al., 2021). Second, weights of an existing sample can be used such that this simulation approximates the actual population. The second method has been applied in this simulation study. The basis of this approach has been developed by Little (1993) and Rubin (1993), discussed by Beckman et al. (1996) and developed in recent applications such as Templ et al. (2017).

In this study, we use the student sample of the German PISA 2018 data as a basis for the simulation (Reiss et al., 2021). By aggregating student data (using school identifiers) to the level of schools, we achieve a school dataset. As the true anonymous list of schools from PISA 2018 with information on the number of PISA eligible students is available to the authors, we add information on the school's MOS, federal state, and school type to the data. We did not only use information from the list of schools because other characteristics, such as student achievement and migration background, are available in the sample.

To simulate the German school frame using a sample of schools, each school has been copied according to its (rounded) school weight. A school from the sample then represents several schools according to their weight in the population. For example, a sampled school with a school weight of 10.21 was copied 10 times on the simulated school frame as it represents about 10 other schools in the population. This approach gives us an approximated copy of the complete school frame. As school weights are adjusted for nonparticipation of schools, this is automatically accounted for in the simulation. Further corrections address changes in the number of fifteen-year-old students between listing and data collection timepoints.

Since students are drawn randomly within schools after sorting by grade and gender, student design weights constitute the inverse of the selection probabilities of students within schools. They are again adjusted for nonresponse of students within schools. Duplicating the students within the schools in the sample by those within school student weights achieves the final simulated population, which can now be used to determine some "true population values", such as mean achievement and its associated standard deviation.

To compare the characteristics of this simulated population with the true school list for Germany in PISA 2018, the total number of students in the frame, the MOS, the federal states, and the school types are used. The true school population comprises 13,855 schools, while the simulated school population cover 13,046 schools. The MOS's mean and standard deviation are slightly higher in the simulated school population (M=58.64, SD=44.58) than in the real population (*M*=52.98, *SD*=43.86). Deviations can be attributed to rounding errors and further sample trimming factors. Rounding errors can be attributed to the rounded school and student weights (to an integer with no decimals) used to create the simulated school and student population. The trimming factors include adjustments when the number of estimated fifteen-year-olds differs significantly from the actual number of those students in a school (there is a period over a year between the listing and testing in a school).

Furthermore, six of the schools drawn did not have any fifteen-year-old students, so that no testing could occur. Two other schools were excluded during the assessment (Mang et al., 2019). Table A1 gives a comprehensive overview of those characteristics.

Analysis Procedures - Stratification, Samples and Weights

Seven different stratification designs have been defined and applied for the simulation study. Table 1 below details the variables used in the different stratification designs under study, whereas Table 3 lists the designs and their explicit, first implicit, and second implicit stratification variables. Additionally, Table 2 details the seven different school types mentioned in Table 1, comprising of lower and upper secondary schools and lower and upper secondary comprehensive schools.

Table 1: Overview of stratification categories and their abbreviations for the simulation study

Abbreviation	**Stratification Categories**	
FS	16 federal states of Germany Special handling of:	
	SAR	Very small federal state Saarland
FS - grouped	CFS	3 city federal states
	NFS	5 "new" federal states
	OFS	8 "old" federal states
MIGRATION	3 Levels of the proportion of students with migration background	
ST	7 School types Special handling of:	
	SEN	Special educational needs
	VOC	Vocational
LOC	3 Levels of competence	

Table 2 – School types used for implicit stratification in the simulation

School type (English translation))	School type (Original name in German)
Lower secondary, some with access to upper secondary; basic general education (exclusively students of the same track)	Hauptschule
Lower secondary, access to upper secondary; extensive general education (exclusively students of the same track)	Realschule
Lower secondary, access to upper secondary; basic and extensive general education	Schule mit mehreren Bildungsgängen
Lower secondary and upper secondary; academic education (exclusively students of the same track)	Gymnasium
Lower and upper secondary comprehensive	Integrierte Gesamtschule
SEN schools	Förderschulen
VOC schools	Berufsschulen

Table 3: Stratification variants for the simulation study

Stratification design	**Explicit stratification**	**Number of explicit strata**	**Implicit stratification**	**Number of implicit strata**	
				Within explicit strata	**Overall**
1	-	1	-	1	1
2	FS (16 states)	18	ST (5 strata)	80	112
	VOC		FS (16 categories)	16	
	SEN		FS (16 categories)	16	
3	FS – grouped (CFS, NFS, OFS)	5	ST (5 strata)	15	21
	VOC		FS – grouped (CFS, NFS, OFS)	3	
	SEN		FS – grouped (CFS, NFS, OFS)	3	
4	MIGRATION (3 levels)	5	ST (5 strata)	15	21
	VOC		MIGRATION (3 levels)	3	
	SEN		MIGRATION (3 levels)	3	
5	ST	7			7
6	ST (7 levels)	8	FS (15 categories)	105	112
	SAARLAND		ST (7 strata)	7	
7	LOC (3 levels)	5	ST (5 strata)	15	21
	VOC		LOC (3 levels)	3	
	SEN		LOC (3 levels)	3	

The different stratification approaches will be described below in detail. According to Baumert et al. (2006), individual characteristics such as gender, migration, grades, socioeconomic status, and school-based characteristics such as school type and grade level are essential predictors of student achievement and hence relevant stratification variables for student assessment surveys. While the PISA within-sampling design is standardized across countries and economies (stratification within schools is done by gender and grades), national variation of school sampling designs is possible. We

hence determined the stratification designs under study accordingly while also considering data availability, as described below.

To get a comprehensive picture of stratification, we use an unstratified sample design (i.e., a simple random sample) as a reference point. This is declared as stratification design *1*.

Stratification design *2* reflects the stratification used in the last cycles of PISA and is, therefore, an essential benchmark for this study. In this design, the explicit stratification is implemented using a two-step process: first, vocational (VOC) and special educational needs (SEN) schools are separated, then, all remaining schools are then separated by federal state. Within the federal-states-strata, schools are sorted by the five school types without VOC and SEN schools. Conversely, all VOC and SEN schools are sorted by federal state. This results in 18 explicit and 112 implicit strata, many of which are very small. This design is the one currently applied in PISA.

Stratification design *3* groups federal states into three categories: city, old, and new federal states. City federal states are the three German cities Berlin, Hamburg, and Bremen, which are politically administered as a state; the distinction between old and new federal states reflects the division of states based on the separation of Germany before the reunification in 1989. Although Germany has been a federal republic since then, major differences exist between the old and new federal states, e.g., in salaries or education structure and curricula (Holtmann, 2020). A potentially better approach would be to merge the federal states based on their mean competencies. However, groups of federal states that are homogenous across all domains do not exist. Another argument against such a division is that it may be difficult to communicate and explain the choice to educational stakeholders. Stratification design *3* addresses the problem of too many small strata in design *2* detailed in the section *Sampling Precision: Sampling Variance* of this paper. In addition, the use of federal states is maintained in a grouped form so that the changes compared to variant *2* are minimal. They can be well defended to lay audiences that may challenge the change of the PISA stratification scheme.

It is well known from numerous PISA analyses that socioeconomic and migration background are significant predictors of student proficiency (OECD, 2019; Sirin, 2005; Stanat & Christensen, 2006). However, recording socioeconomic background is difficult, especially in Germany, as this is subject to strict data protection regulations. However, one piece of information available for German schools is the percentage of students with an immigrant background. Therefore, we decided to define stratification design *4* based on these properties. This variant uses categories of schools with different proportions of students with a migration background. Schools having no students with migration background are allocated to the first category of this index. Categories two and three are defined in Table 4 as schools with more than 0% and less than 30% of students with migration background and schools with more than 30%, respectively.

Stratification designs *5* and *6* address school types as explicit stratification variables. For variant *6,* an additional explicit stratum for the federal state of Saarland is created. This is to avoid sampling no schools from this (very small) federal state, which could happen by chance because the number of students in this state is smaller than the sampling interval[2]. Note that including no schools from Saarland in the sample is politically sensitive, and hence, should be avoided. As for the special handling of VOC and SEN schools, the explicit stratification is formed using two steps for this variant: first, the schools from the federal state Saarland (SAR) are separated, and all remaining schools are then separated by school type. Within the school types, schools are sorted implicitly by federal states. Conversely, all schools in SAR are sorted implicitly by their school types.

Table 4: Level of competence for the three domains reading, science, and math; derived competence levels for stratification/Thresholds for the migration index for stratification

Variable	**Thresholds**		
MIGRATION	MIGRATION=0	0<MIGRATION<30	MIGRATION>=30
LOC	LOC<=400	400<LOC<500	LOC>=500

[2] The sampling interval is the sum of the number of fifteen-year-olds in all schools divided by the number of schools to be sampled in each stratum.

Stratification design *7* from Table 3 represents the near-optimal stratification variant, where an aggregated index of student competence is used to categorize schools into three performance levels. The LOC is not available for German schools with official statistics and therefore is used just as another benchmark design in this study. It is defined in this study based on the 10 plausible values (PVs) of the three main domains of math, reading, and science obtained in PISA 2018 (OECD, 2020); these were combined at the individual student level and then aggregated to the school level. Each school was allocated to one of the three categories in Table 4. PVs, representing the competency of one student, are 10 drawn values from the answering distribution of this pupil to the PISA testing questions. The answering distribution is based on the principles of Item Response Theory (IRT; Rasch, 1960) and adapted to PISA actual standards by Davier and Sinharay (2013). With IRT models, student responses to the questions from the PISA test are modelled as a probability function of person and item characteristics. For example, detailed explanations of this estimation procedure can be found in OECD (2020) and Mang et al. (2019).

Note that VOC schools and SEN schools are treated as separate strata in stratification variants *3, 4,* and *7* because students in these school types perform systematically lower than students in other school types. Separation further allows for achieving higher precision for these groups of students by oversampling schools in these strata. Additionally, the implicit sorting by school type is retained for variants *3, 4,* and *7* as it is highly related to achievement and, therefore, essential for low sampling variance. This sorting also accommodates a higher precision for comparisons between school types.

For each stratification design, the frame is sorted by explicit and then implicit stratification and then by MOS in a serpentine manner, mimicking the PISA sorting method. In the next step, 2,000 samples of 223 schools with 30 students per school were drawn by systematic PPS sampling for each stratification scenario using a Monte Carlo approach. The sample size of 223 schools and 30 students per school was chosen, as this number reflects the number of schools and students participating in PISA 2018 in Germany. Please note the standard PISA sampling international target is 150 schools and 42 students per school (OECD, 2020). The PPS sampling procedure implies that schools are

selected using a random start and a sampling interval within the explicit strata. Schools are selected for the sample if the cumulative sampling interval matches the cumulative number of fifteen-year-olds in the schools.

Within the schools, an equal probability sample of PISA students was selected using systematic sampling, where the lists of students were first sorted by grade and then by gender. In schools with less than 30 eligible students, all of them were selected. Using the binomial distribution or so-called Bernoulli processes (Clopper & Pearson, 1934), it is determined that 2,000 replicates are adequate to achieve a coverage probability of greater than 99% for the 95% confidence interval of the estimates. This approach allowed a nearly exact representation of the sampling distribution, thereby enabling a precise estimation of the sampling precision (i.e., the SEs of specific population features) for each scenario.

The school and student base weights were automatically generated after drawing the school and the student sample for each stratification variant. Therefore, the estimation weight we use in our simulation is the product of the school and the student base weight, given by

$$w_{ij} = \frac{1}{\pi_{ij}}, \tag{4}$$

with π_{ij}= selection probability for student *j* given school i has been selected. The calculation of replicate weights to correctly estimate the SEs in this study is based on the BRR method with Fay's adjustment (Judkins, 1990), as done in PISA (OECD, 2009, 2020). Preserving the order of schools in the sample determined by the sorting before the selection process, two adjacent schools belonging to an explicit stratum are paired into so-called variance zones. If there is an odd number of schools in a stratum, the last group is set with three schools. Once 80 variance zones are reached, the next pair of schools is again allocated to zone one, the second-to-next pair to zone two, and so on. One school within the pairs is randomly numbered as one, the other as two. In the case of three schools being placed in a zone, one school is randomly numbered as one and the other two schools as two. With

the help of these variance zones, 80 replicate weights are then calculated with the help of a Hadamard matrix explained in the section *Estimation procedures for multistage, stratified PPS sampling: Variance estimation* in this paper.

Nonresponse for both levels must also be considered to determine the final school and student weights. As the assessment is mandatory in Germany, nonresponse for schools was very low over most cycles. Hence, we assumed 100% participation at the school level for the simulation. Furthermore, student nonresponse is not the focus of this article and is therefore also neglected (100% student participation is assumed). Some minor adjustments to student base weights regarding, e.g., school nonparticipation or corrections from the estimation of the number of fifteen-year-olds were applied to reflect the true population values as precisely as possible in the samples.

One constraint of this simulation study is that measurement variance might be underestimated as one student in the base sample with a given competency represents multiple students with exactly this competency value (represented by PV's) in the simulated population. That is, a student with, say, a competency score of 500 and a total student weight of 200 represents 200 other students with the same competency of 500 and, thus, no variation among those students. To account for this simulation feature, random noise is added to each of the 10 PVs of the individual domains. This is added to the original PVs via random selection from a normal distribution with a mean of 0 and a 1/4 fraction of the standard deviation of the respective PVs grouped by school type. This proportion was chosen based on evidence, as it adds "noise" to the distribution of skills without changing the distribution characteristics.

Given the stratification designs used with the 2,000 samples and associated weights and replicate weights, mean calculations for the three PISA domains reading, science, and math and their associated SEs were calculated in the following step, and these 2,000 estimates per variant and domain were compared with their distributional properties in the following sections.

Results and Discussion

Variance in Proficiency Explained by Stratification

As explained earlier in this paper, efficient stratification variables are closely related to the outcome variables. Therefore, using a regression modelling approach, we examined in a first step what part of the variance of the achievement scores was explained by the stratification variants, implicit and explicit stratification, in the different scenarios (Table 5 and Table 6*).*[3] Table 5 shows the variances of average school proficiency explained by the stratification scheme in each design, whereas Table 6 displays the respective variances in student proficiency. Average school proficiency was determined by the average student proficiency for each subject, using the first plausible value for each student. Note that only the first PV for mathematics, science, and reading was used as it approximates the distribution of student achievement correctly (Davier et al., 2009).

Table 5: Explained variances in average school proficiency (math, science, and reading) by explicit and implicit stratification for each stratification variant. Method: linear regression

Stratification design	Explicit Stratification	Implicit Stratification	Mathematics	Science	Reading
1	-	-	-	-	-
2	FS, VOC, SEN	ST, FS	0.86	0.84	0.84
3	FS – grouped, VOC, SEN	ST, FS	0.83	0.81	0.81
4	MIGRATION, VOC, SEN	ST, MIGRATION	0.82	0.80	0.80
5	ST	-	0.82	0.79	0.79
6	ST, SAR	FS, ST	0.86	0.84	0.83
7	LOC, VOC, SEN	ST, LOC	0.91	0.89	0.89

[3] The OECD (2020) also displays information on explained variances (Annex C6). Note that they are based on multilevel models, drawing on information from both students and schools simultaneously, and can therefore not be compared with the information presented in Tables 5 and 6.

Table 6: Explained variances in student proficiency (math, science, and reading) by explicit and implicit stratification for each stratification variant. Method: linear regression

Stratification design	Explicit Stratification	Implicit Stratification	Mathematics	Science	Reading
1	-	-	-	-	-
2	FS, VOC, SEN	ST, FS	0.38	0.38	0.39
3	FS – grouped, VOC, SEN	ST, FS	0.36	0.36	0.38
4	MIGRATION, VOC, SEN	ST, MIGRATION	0.35	0.36	0.37
5	ST	-	0.33	0.34	0.36
6	ST, SAR	FS, ST	0.37	0.37	0.38
7	LOC, VOC, SEN	ST, LOC	0.66	0.69	0.72

Comparing Tables 5 and 6, the first thing to emphasize is that differences across schools explain about half of the variance in student proficiency (variances from Table 5 are about double compared to those from Table 6), meaning that the school context can explain a large proportion of the explained variance. Stratification design *1* represents the variant without stratification. Hence, no variance can be explained by this scheme. Variants *2 to 6* explain about one-third of the variance in students' proficiency scores in the three competencies math, science, and reading (Table 6). Stratification variant *6* slightly outperforms variants *3*, *4,* and *5*, explaining the same variance as variant *2*. As expected, the near-perfect stratification variant *7* illustrates the highest share in proficiency score variance since it is based on the proficiency scores themselves. In addition to these findings, we calculated these explained variances using the "actual" data from the PISA 2018 sample and presented them in Tables A3 and A4 in the appendix. It can also be seen that, for the sample, the variance explanations at the school level are almost twice as high as at the student level. Also, the proportions of explained variance at the school level compared to the simulated population values are almost identical, with a bias of approximately three to four percentage points found at the student level. This may be due to the fixed stratification in PISA 2018 with stratification design *2,* or it may be due to the added "noise" to the PVs (please refer to section *Analysis Procedures - Stratification, Samples and Weights* for the explanation). In summary, this analysis can serve as a

basis and interpretation aid for the simulation study results. It provides the first evidence that stratification variants *3, 4, 5, and 6* can likely be a reasonable alternative to the currently implemented variant *(2)*.

Results of the Simulation Study

We present further results in the format of boxplots and tables. Boxplots describe the distribution of the estimated values each based on many repetitions (2,000 in our study). The median, the 25th, and 75th percentiles, minimum and maximum, are presented (Chambers, 1983). Differences between the boxplots are interpreted based on several definitions (e.g., Williamson et al., 1989). First, the boxes representing the interquartile ranges are compared. If boxes do not overlap, a difference can be stated. Second, 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. Third, 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 a wider distribution, that is, more scattered data.

In Tables 7, 8, and 9, informal statistics are listed for math, science, and reading for all seven stratification designs. Augmenting the upcoming graphical results in Figure 1 and Figure 2, the tables provide the following information. Column 1 (Mean math bias) presents the deviation from the estimated mean of the respective competences to the true mean values of the population. Column 2 shows each parameter's empirical 95% coverage rates (CR). The empirical 95% coverage rate indicates how often each estimated parameter's 95% confidence interval covers the true population value. An acceptable coverage rate starts at 95%. Column 3 presents the SEs computed using the BRR method, averaged over the 2,000 sample replicates. Column 4 displays the "true" SE for each variant, calculated as the standard deviation (SD) of the average student achievement over the 2,000 sample replicates, i.e., the SD of the sampling distribution. Finally, we present in column 5 Root Mean Squared Error (RMSE). A low RMSE value means that the estimator's bias and variance are small.

Table 7: Mean bias, SEs, and fit statistics for the domain math by stratification design.

Stratification design	(1) Mean Math Bias	(2) CR Mean Math	(3) Mean Math SE (BRR)	(4) Mean Math SE (SD of sampling distribution)	(5) RMSE
1	0.02	0.98	4.28	4.12	4.12
2	0.03	1.00	3.00	2.48	2.48
3	0.04	1.00	2.71	2.41	2.41
4	0.50	1.00	2.34	1.45	1.53
5	0.12	1.00	2.22	2.39	2.39
6	0.04	1.00	2.16	2.51	2.51
7	0.37	1.00	1.58	1.30	1.35

Note: SE=sampling error, CR=coverage rate, BRR= balanced repeated replication, RMSE=root mean squared error

Table 8: Mean bias, SEs, and fit statistics for the domain science by stratification design.

Stratification design	(1) Mean Science Bias	(2) CR Mean Science	(3) Mean Science SE (BRR)	(4) Mean Math SE (SD of sampling distribution)	(5) RMSE
1	0.23	0.97	4.46	4.09	4.10
2	0.29	1.00	3.06	1.92	1.95
3	0.28	1.00	2.78	1.82	1.84
4	0.59	1.00	2.45	1.35	1.48
5	0.35	1.00	2.31	1.83	1.86
6	0.25	1.00	2.21	1.92	1.94
7	0.41	1.00	1.66	1.24	1.30

Note: SE=sampling error, CR=coverage rate, BRR= balanced repeated replication, RMSE=root mean squared error

Table 9: Mean bias, SEs, and fit statistics for the domain reading by stratification design.

Stratification design	(1) Mean Reading Bias	(2) CR Mean Reading	(3) Mean Reading SE (BRR)	(4) Mean Reading SE (SD of sampling distribution)	(5) RMSE
1	0.72	0.97	4.80	4.73	4.78
2	0.64	1.00	3.45	2.92	2.99
3	0.70	1.00	3.12	2.83	2.91
4	0.08	1.00	2.62	1.54	1.54
5	0.58	1.00	2.74	2.77	2.83
6	0.71	1.00	2.79	2.98	3.06
7	0.30	1.00	2.11	1.42	1.45

Note: SE=sampling error, CR=coverage rate, BRR= balanced repeated replication, RMSE=root mean squared error

Although stratification only impacts the estimation precision, columns 1 and 2 of Tables 7, 8, and 9 show that all methods estimate the mean domain value with little bias, as expected.

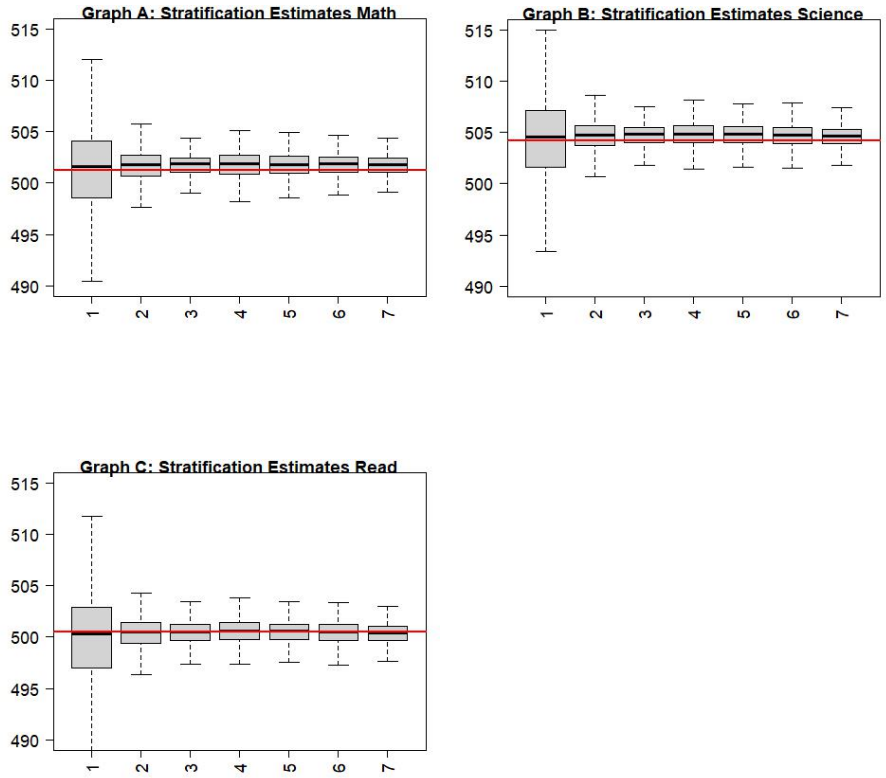

Figure 1: Distribution of estimates for proficency means by stratification variant; please refer to table 3 for the description of the stratification variants.

Figure 1 augments and confirms the information presented in the tables, with boxplot panels A to C presenting the distribution of estimated means for the three proficiency domains based on the simulation (2,000 samples). The red horizontal line represents the true population value. Looking closely at Figure 1, we see that the true values are optimally covered for the analysed domain reading (graph C). At the same time, a consistent but negligibly slight bias appears for the estimation of means for mathematics and science.

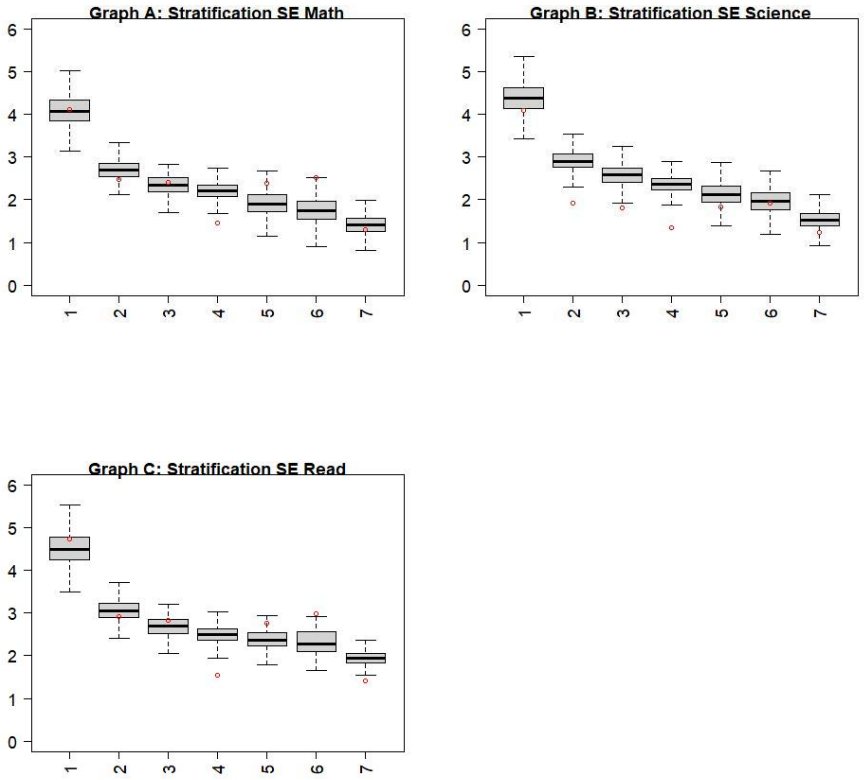

Figure 2: Distribution of estimated sampling error with BRR by stratification variant; please refer to table 3 for the description of the stratification variants.

This research focuses on sampling precision, which is presented in columns 3 and 4 of Tables 7 to 9, augmented by a graphical display (Figure 2) of the distributions of SE estimates of the domain means, here based on BRR.[4] The red points in the figure indicate the "true" SE measured by the standard deviation of the sampling distribution. The findings are equivalent for all domains. As expected,

[4] Please consider that deviations from SEs displayed in Table 12.7 in the PISA 2018 Technical Report (OECD, 2020) and reported SEs (BRR) in Table 7,8,9 are due to the simulation design.

stratification design *1* (i.e., unstratified sample) results in the highest SEs and stratification design *7* (i.e., stratification by average proficiency) results in the smallest SEs. The remarkable difference shows the potential of optimal stratification: comparing designs *1* and *7*, SEs decrease by a factor of three, equivalent to an increase in sample size by roughly a factor of 10, given no changes in the sampling design. In other words, if one wishes to decrease sampling precision by the same factor without changing the stratification design, one must select a sample that is ten times bigger.

The stratification design that PISA currently applies (stratification design *2*) decreases SE, too, on average, across the three domains by a factor of around 1.5 compared to no stratification. This is equivalent to doubling the sample size. However, stratification designs *3* to *6* all outperform design *2*. SEs are almost halved compared to design *1* (no stratification), equivalent to an increase in sample size by a factor of three. Designs *5* and *6* show the best results regarding sampling precision. However, the gains are minimal compared to designs *3* and *4*. However, looking strictly at the true SE (column 4 in Tables 7 to 9), only design 4 results in substantially smaller SEs than design *2*.

Another side effect of stratification is that the precision of the SE estimates is higher – this can be seen in Figure 2. The distances between the boxplot whiskers are smaller in all variants applying stratification.

By looking at RMSE, we account for both sampling precision and proficiency estimation accuracy (columns 5 in Tables 7 to 9). Again, unsurprisingly, the highest and lowest RMSE is observed in variants *1* and *7*, respectively. RMSE values are similar for variants *2*, *3*, *5*, and *6*, while variant *4* shows the best performance again.

Biasedness of Sampling Error Estimates when using BRR

A rather unexpected finding of this simulation study was the discrepancy in the SEs when comparing the "true" values (computed as the SD of the sampling distribution over 2,000 samples) versus the averaged SEs estimated using BRR. This is not the focus of this paper but warrants further investigation, which is why we briefly describe the issue in this section. The BRR SE estimates are – with a few exceptions - consistently larger than the true values. That means the standard errors seem

to be systematically overestimated. After careful consideration, there was a presumption that estimating standard errors using BRR does not comprehensively account for implicit stratification. The authors re-performed all analyses with a random permutation for the applied implicit stratification variables to address this hypothesis. Unlike the implicit stratification in stratification designs *2-7*, there is now a random implicit sorting assuming no implicit sorting was applied. In doing so, the standard deviations of the estimates (i.e., the "true" SE) become visibly larger and approach the true SEs (see Table A2 with 100 replicates in the appendix). Without implicit sorting, different (i.e., less precise) mean estimators result for each sample so that the overall sampling distribution has a larger standard deviation. Note that we present only analysis for the domain math in the Appendix; for the other domains, the outcomes are comparable.

Related to this, note that the coverage rates presented in columns 2 of Tables 7 to 9 above were estimated based on the BRR SEs. It can be seen that almost all stratification designs achieve 100% CR meaning that all true population values were covered in the 95% confidence interval of each estimated parameter. Given the results above, it can be assumed that the CR is overestimated.

Furthermore, some approaches note and discuss an overestimation of the standard deviation from the sampling distribution via replication approaches such as BRR or similar methods, e.g., the JRR method (Qian, 2020; Rizzo & Judkins, 2004; Rizzo & Rust, 2011). Other variants for estimating the standard error based on Taylor series expansion (Lavrakas, 2008; Valliant et al., 2018b), such as the so-called delta method (Cochran, 1977), seem to result in more robust and efficient estimates (Krewski & Rao, 1981; Qian, 2020; Wolter, 2007). A problem of this variant is that it requires the joint inclusion probability for each variance zone, i.e., the probability that the two selected schools in the respective variance zone are jointly selected. This probability can become zero for certain pairs of units within the chosen variance estimation process (Wolter, 2007). However, there are ways to estimate it (Hajek, 1964; Särndal et al., 2003). To consider all confounding parameters of this discrepancy in the simulation, parts of the simulation were also calculated with JRR. An almost identical result structure confirms the suspicion of the conservative estimation of the standard

deviation of the estimated values by repeated replication methods. In addition, it should be mentioned that the "ideal" conditions of the simulation study probably also underestimate the SD of the sampling distribution since specific "errors" such as schools' or students' nonresponse may not be considered.

Summary and Conclusions

This simulation study reflects the relevance of stratification and, in particular, its high potential for efficient sample designs in the case of PISA Germany.

First, the study reconfirmed that stratification does not affect parameter estimation, here looking at the mean achievement of the PISA domains mathematics, science, and reading. More importantly, we found large differences in the SEs of achievement scores when applying different stratification schemes for school sampling. This study aimed to investigate alternative stratification designs since the one currently applied results in strata that are too small, causing technical problems when preparing the sample data for inference statistics (i.e., estimation of population features). One problem is that the small strata cause suboptimal data handling for estimating sampling variance with BRR. Explicit strata had to be collapsed in previous cycles to accommodate the pairing algorithm in BRR, a procedure that compromises technical standards. Further, nonresponse adjustment procedures were affected (an issue not covered in this article).

We studied four alternative stratification designs, referred to as designs *3*, *4*, *5*, and *6*, that all overcome the problem of small strata, and compared them with the current scheme (design *2*), a variant without stratification (design *1*), and an optimal stratification design (*7*).

Considering the true SEs and the RSME exclusively, design 4 performs best. However, switching to this stratification design would lead to a substantial change in the PISA sampling design. This scheme stratifies based on the proportions of students with a migration background and completely neglects the German school structure tied to federal states. This would change the logistics for conducting the PISA study in Germany, as it would, for example, be impossible to allocate a fixed number of schools

to each federal state and inform states at an early stage about sample sizes to be expected. It is also possible that no schools at all are drawn from very small states (especially the Saarland). Given these effects, stratification design *4* may not be the best solution for a change. Note this is not a problem from a methodological point of view: no comparisons between federal states are intended for PISA, and the sample remains unbiased.

Designs *3*, *5*, and *6* can also be recommended as alternatives. They show sufficiently good estimate precision and BRR SEs are smaller than variant *2*.

Stratification design *3* groups the federal states into three categories (city states, old and new German states). Since this grouping preserves the federal-state structure of Germany, it may provide one good stratification design alternative for upcoming cycles of the PISA study, representing a conservative and cautious change. However, it does not entirely overcome the logistical issues pointed out above for design *4*. By an implicit stratification by federal states (designs *5* and *6*), the issue of unpredictable sample sizes can be solved, as this procedure results in a close-to-perfect proportional allocation of the sample to all strata so that the sample sizes per federal state become predictable. Variant *6 also* solves the issue of the likelihood of selecting no school in Saarland. Both designs *5* and *6* use the types of schools for explicit stratification, ensuring high sampling precision as school type is very closely related to the average proficiency of students. Overall, we believe that stratification design *6* meets all requirements of a stratification design in Germany and can therefore be thoroughly recommended for future PISA cycles.

The reduced SEs with a change in stratification will lead to more precise samples, smaller confidence intervals, and higher statistical power when comparing Germany with other participating countries, economies, or specific groups of students within Germany (e.g., gender differences). Increased statistical power may allow the comparison of smaller subgroups, which was not possible before. However, this may involve communication challenges, i.e., explaining specific findings to a lay audience. For example, a difference of 5 points between two comparison groups would not have

been detected as a statistically significant difference in previous cycles, but now would. While a statistician is aware that an insignificant result does not mean there is *no* difference between groups but merely means we cannot know whether or not there is a difference, this is a misinterpretation that is very common even among scholars less familiar with statistical theory. In connection with trend calculations between two PISA cycles and their cross-sectional nature, it can be stated that the linking error, considering the uncertainty between two assessments, might increase due to the proposed change in the sampling design (OECD, 2020). The complete SE consisting of sampling, imputation, and linking error will then increase, and results might not become as statistically significant as they would without changing the sampling design.

Suppose an increase in sampling precision is not needed or not desired. In that case, another possibility is a change in the stratification design and a reduction in sample size while keeping precision constant with previous cycles. This could reduce the burden on German schools that must cope with various regional, national, and international studies and assessments. This could also mean that resources are directed toward better data quality rather than "more data." For example, a smaller sample size means national centres can direct funds to increase participation rates. In any case, a change in the stratification design for PISA in Germany must be carefully communicated with relevant stakeholders (for example, the press or teacher unions) and policymakers.

Future research and initiatives may focus on further possibilities to increase sampling efficiency without increasing costs (Biemer & Lyberg, 2003; Groves, 2011). One direction could be to consider including better socioeconomic background indicators of the student intakes of schools in the sampling frame and the stratification scheme since this is a powerful predictor of student achievement in the PISA domains of mathematics, science, and reading. Another, perhaps even more straightforward, approach would be to use achievement indicators for schools, i.e., categorizing schools by the average achievement of their students. Such indicators could be based on regional mandatory census assessments. As shown with stratification variant *7,* this would be the most efficient design. This approach is already used for several countries in many contemporary large-

scale assessments (e.g., Mullis et al., 2016). While this data also exists in Germany, it is inaccessible for the teams preparing the German school sampling frames for national and international large-scale assessments because of its confidential nature. Providing this data to these teams while adhering to strict data protection measures would be desirable.

Limitations and Outlook

It should be noted that this simulation study has been conducted under ideal conditions. As mentioned earlier in the report, no bias due to nonparticipation was considered at both the school and student levels. Further, even if unlikely, new strategies may also increase other sources of error, or new biases may arise. We refer to the theory of the total survey error (Assael & Keon, 1982; Weisberg, 2005), which introduces non-sampling error sources, such as errors due to frame construction, the sample selection process, data collection, data processing, and estimation methods.

Another limitation of the study is that the proportion of foreign students in schools, which is used as a stratification design in *Stratification 4*, does not consider whether a student with an immigrant background has a German passport because, unlike their parents, they were born in Germany. Since public statistics are usually not allowed to publish these subtleties due to data protection, this aspect must be taken care of in the stratification for interpretations. Another limitation here may be that this information may not be consistently available in public statistics the frame is based on, and hence, the effect might be overestimated. Furthermore, it would be desirable to calculate additional statistics, such as correlation or regression coefficients, to quantify the precision gain further. Finally, the discrepancy between the true SEs and their estimation via BRR should be examined in more depth. In particular, the relationship between BRR and the origin of Taylor Series Linearization (Lavrakas, 2008; Valliant et al., 2018b) with its application of the delta method (Cochran, 1977) shall be addressed in future studies.

Last but not least, our results are hardly transferable to other studies as explicitly only the stratification of Germany in PISA has been addressed. However, it may serve as a guide for other

countries establishing or revising their stratification. It should be considered that proportions in the school or student population might change and need to be considered in future adjustments. So can migrational movement lead to changed population characteristics that must be controlled to apply the given suggestions.

In summary, it can be emphasized that the principle of stratification with its systematic sampling should be retained in the complex sampling design in PISA, but with recommended adjustments in the execution of explicit and implicit execution of stratification.

List of abbreviations

BRR	Balanced Repeated Replication
CI	Confidence Interval
HT	Horvitz Thompson
IDB	International Database (Analyzer)
ICC	Intraclass Correlation
IEA	International Association for the Evaluation of Educational Achievement
IRT	Item Response Theory
JRR	Jackknife Repeated Replication
LSA	Large-Scale Assessment
MOS	Measure of Size
MSE	Mean Squared Error
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PPS	Probability Proportional to Size
PSU	Primary Sampling Units
SE	Standard Error

SD	Standard Deviation
SRS	Simple Random Sample
TIMSS	Trends in International Mathematics and Science Study

Appendix

*Table A1: Comparison of the simulated school population and the true school population of PISA 2018 (Frame). Due to data protection reasons, strata were pseudonymized**.*

	Simulated school population	**PISA 2018 school population (Frame)**
N	13046	13855
Mean MOS	58.64	52.98
SD MOS	44.58	43.86
N FS 1	1995	2002
N FS 2	2187	1913
N FS 3	229	299
N FS 4	263	295
N FS 5	60	80
N FS 6	183	163
N FS 7	829	920
N FS 8	237	291
N FS 9	1218	1279
N FS 10	1889	2068
N FS 11	401	417
N FS 12	93	103
N FS 13	515	516
N FS 14	247	306
N FS 15	303	428
N FS 16	335	374
N SEN	1208	1334
N VOC	854	1067
N ST 1	2857	2559
N ST 2	1915	2077
N ST 3	1514	1742

N ST 4	3102	3129
N ST 5	1596	1947
N ST 6	-*	
N ST 7	2062	2401

**No school type 6 has been sampled in PISA 2018.*
***The MOS, the explicit stratification of PISA 2018 (FS: federal states, special educational needs, and vocational schools), the suggested grouped explicit stratification (FS new), and the school types are displayed. The absolute number, means, and standard deviations have been analysed.*

Table A2: Mean bias, SEs, and fit statistics for the domain math by stratification design with a random permutation per sample for the applied implicit stratification variables with 100 replications

Stratification variant	**(1) Mean Math Bias**	**(2) CR Mean Math**	**(3) Mean Math SE (BRR)**	**(4) Mean Math SE (SD of sampling distribution)**	**(5) RMSE**
1	0.14	0.97	4.27	4.45	4.43
2	0.10	0.96	4.20	4.22	4.20
3	0.28	0.99	4.19	4.16	4.15
4	0.01	0.95	3.98	4.28	4.26
5**	-	-	-	-	-
6	0.40	0.97	2.38	2.99	3.00
7	0.20	0.97	2.15	2.70	2.69

Note: SE=sampling error, CR=coverage rate, BRR= balanced repeated replication, RMSE=root mean squared error
***no implicit stratification for* Stratification 5 *variant*

Table A3: Explained variances in average school proficiency (math, science, and reading) by explicit and implicit stratification for each stratification variant for real PISA 2018 sample data. Method: linear regression

Stratification design	Explicit Stratification	Implicit Stratification	Mathematics	Science	Reading
1	-	-	-	-	-
2	FS, VOC, SEN	ST, FS	0.87	0.86	0.86
3	FS – grouped, VOC, SEN	ST, FS	0.82	0.83	0.83
4	MIGRATION, VOC, SEN	ST, MIGRATION	0.81	0.81	0.81
5	ST	-	0.80	0.80	0.81
6	ST, SAR	FS, ST	0.86	0.86	0.85
7	LOC, VOC, SEN	ST, LOC	0.90	0.90	0.90

Table A4: Explained variances in student proficiency (math, science, and reading) by explicit and implicit stratification for each stratification variant for real PISA 2018 sample data. Method: linear regression

Stratification design	Explicit Stratification	Implicit Stratification	Mathematics	Science	Reading
1	-	-	-	-	-
2	FS, VOC, SEN	ST, FS	0.42	0.41	0.43
3	FS – grouped, VOC, SEN	ST, FS	0.39	0.39	0.41
4	MIGRATION, VOC, SEN	ST, MIGRATION	0.39	0.39	0.41
5	ST	-	0.38	0.38	0.40
6	ST, SAR	FS, ST	0.41	0.41	0.43
7	LOC, VOC, SEN	ST, LOC	0.69	0.73	0.75

References

Assael, H., & Keon, J. (1982). Nonsampling vs. Sampling Errors in Survey Research. *Journal of Marketing*, *46*(2), 114. https://doi.org/10.2307/3203346

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). *4 Sampling designs of the National Educational Panel Study: challenges and solutions* (Vol. 14). https://doi.org/10.1007/s11618-011-0181-8

Baumert, J., Stanat, P., & Watermann, R. (Eds.). (2006). *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit: Vertiefende Analysen im Rahmen von PISA 2000* (1. Aufl.). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90082-7

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part a: Policy and Practice*, *30*(6), 415–429. https://doi.org/10.1016/0965-8564(96)00004-3

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. *Wiley series in survey methodology*. Wiley-Interscience. https://doi.org/10.1002/0471458740

Brown, R. S. (2010). Sampling. In Peterson P. L., E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (S. 142–146). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-044894-7.00294-3

Buchmann, C., & Park, H. (2009). Stratification and the formation of expectations in highly differentiated educational systems. *Research in Social Stratification and Mobility*, *27*(4), 245–267. https://doi.org/10.1016/j.rssm.2009.10.003

Chambers, J. M. (1983). *Graphical methods for data analysis*. *Chapman & Hall statistics series*. Wadsworth & Brooks/Cole.

Clopper, C. J., & Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, *26*(4), 404. https://doi.org/10.2307/2331986

Cochran, W. G. (1977). *Sampling techniques* (3. ed.). *A Wiley publication in applied statistics*. Wiley.

Davier, M. von, 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.*

Davier, M. von, & Sinharay, S. (2013). Analytics in International Large-Scale Assessments: Item Response Theory and Population Models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment.* Chapmall Hall/CRC; Chapman and Hall/CRC.

Groves, R. M. (2011). *Survey Methodology* (2nd ed (Online-Ausg.)). *EBL-Schweitzer: v.561*. Wiley.

Hajek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. *The Annals of Mathematical Statistics*, *35*(4), 1491–1523. https://doi.org/10.1214/aoms/1177700375

Hansen, M. H., & Hurwitz, W. N. (1943). On the Theory of Sampling from Finite Populations. *The Annals of Mathematical Statistics*, *14*(4), 333–362. https://doi.org/10.1214/aoms/1177731356

Holtmann, E. (2020). Deutschland 2020: unheilbar gespalten? *Zeitschrift für Politikwissenschaft*, *30*(3), 493–499. https://doi.org/10.1007/s41358-020-00223-6

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. https://doi.org/10.2307/2280784

Jaeger, R. M. (1984). *Sampling in education and the social sciences*. Longman.

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*(No. 6), 223–239.

Kauermann, G., & Küchenhoff, H. (2011). *Stichproben: Methoden und praktische Umsetzung in R*. Springer; Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12318-4

Kish, L. (1965). *Survey Sampling.* John Wiley and Sons.

Krewski, D., & Rao, J. N. K. (1981). Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods. *The Annals of Statistics*, *9*(5), 1010–1019. http://www.jstor.org/stable/2240615

Lavrakas, P. (2008). Taylor Series Linearization (TSL). In P. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods - Volume 1.* Sage Publications. https://doi.org/10.4135/9781412963947.n572

Levy, P. S., & Lemeshow, S. (2013). *Sampling of Populations: Methods and Applications*. John Wiley & Sons.

Little, R. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, *Vol. 9 No. 2*, 407–426.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, *9*(8). https://doi.org/10.18637/jss.v009.i08

Mahalanobis, P. C. (1952). Some aspects of the design of sample surveys. *The Indian Journal of Statistics*(12), 1–7.

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education*, *9*(1), 1–39. https://doi.org/10.1186/s40536-021-00099-0

Mang, J., Wagner, S., Gomolka, J., Schäfer, A., Meinck, S., & Reiss, K. (2019). *Technische Hintergrundinformationen PISA 2018*. Technische Universität München. https://doi.org/10.14459/2019MD1518258

Meinck, S. (2020). Sampling, Weighting, and Variance Estimation. In H. Wagemaker (Ed.), *IEA Research for Education, A Series of In-depth Analyses Based on Data of the International Association for the Evaluation of Educational Achievement (IEA). Reliability and Validity of International Large-Scale Assessment: Understanding IEA's Comparative Studies of Student Achievement* (1st ed., pp. 113–129). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_7

Meinck, S., & Vandenplas, C. (2021). Sampling Design in ILSA. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings* (pp. 1–25). Springer International Publishing. https://doi.org/10.1007/978-3-030-38298-8_25-1

Mullis, I. V., Martin, M. O., & Hooper, M. (2016). *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College.

OECD. (2002). *PISA 2000 Technical Report*. Organisation for economic co-operation and development. https://doi.org/10.1787/9789264199521-en

OECD. (2009). *PISA data analysis manual: SPSS* (Second edition). OECD. https://doi.org/10.1787/9789264056275-en

OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing.

OECD. (2019). *PISA 2018 Results (Volume II): Where all students can succeed.* OECD. https://doi.org/10.1787/b5fd1b8f-en

OECD. (2020). *PISA 2018 Technical Report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Qian, J. (2020). Variance Estimation with Complex Data and Finite Population Correction--A Paradigm for Comparing Jackknife and Formula-Based Methods for Variance Estimation. Research Report. Ets RR-20-11. *ETS Research Report Series*.

R Core Team. (2020). *R: A language and environment for statistical computing [*Computer software]. R Foundation for Statistical Computing. Vienna, Austria.

Rao, J. N. K., & Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, *86*(2), 403–415. https://doi.org/10.1093/biomet/86.2.403

Rao, J. N. K., & Wu, C. F. J. (1985). Inference From Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics. *Journal of the American Statistical Association*, *80*(391), 620. https://doi.org/10.2307/2288478

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks pædagogiske Institut.

Reiss, K., Mang, J., Heine, J.-H., Weis, M., Schiepe-Tiska, A., Diedrich, J., Klieme, E., & Köller O. (2021). *Programme for International Student Assessment 2018 (PISA 2018). Dataset*. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_PISA_2018_v1

Rizzo, L., & Judkins, D. R. (2004). Replicate Variance Estimation for the National Survey of Parents and Youths. *JSM Proceedings: Survey Research Method Section*, 4257–4263.

Rizzo, L., & Rust, K. F. (2011). Finite Population Correction (FPC) for NAEP Variance Estimation. *JSM Proceedings: Survey Research Method Section*, 2501–2515.

RStudio Team. (2020). *RStudio: Integrated Development Environment for R* (Version 1.4.1717) [Computer software]. RStudio, Inc. Boston, MA.

Rubin, D. B. (1993). Discussion statistical Disclosure Limitation. *Journal of Official Statistics*(Vol.9 No. 2), 461–468.

Rust, K. F., & Rao, J. N. K. (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., Davier, M. von, & Rutkowski, D. (Eds.). (2013). *Handbook of International Large-Scale Assessment*. Chapmall Hall/CRC; Chapman and Hall/CRC. https://doi.org/10.1201/b16061

Särndal, C.-E., Swensson, B., & Wretman, J. H. (2003). *Model assisted survey sampling*. *Springer series in statistics*. Springer.

Singh, R., & Mangat, N. S. (1996). *Elements of Survey Sampling* (Vol. 15). Springer-Science+Business Media, B.V.; Springer Netherlands. https://doi.org/10.1007/978-94-017-1404-4

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Skinner, C. J. (2014). Probability Proportional to Size (PPS) Sampling. *Wiley StatsRef: Statistical Reference Online*, 1–5. https://doi.org/10.1002/9781118445112.stat03346.pub2

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What Are We Weighting For? *Journal of Human Resources*, *50*(2), 301–316. https://doi.org/10.3368/jhr.50.2.301

Stanat, P., & Christensen, G. S. (2006). *Where Immigrant Students Succeed: A Comparative Review of Performance and Engagement in PISA 2003*. *Programme for International Student Assessment (PISA)*. OECD.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., & Asok, C. (Eds.). (1984). *Sampling Theory of Surveys with Applications*. Iowa State University Press; Ames and Indian Society of Agricultural Statistics.

Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of Synthetic Complex Data: The R Package simPop. *Journal of Statistical Software*, *79*(10). https://doi.org/10.18637/jss.v079.i10

Thompson, S. K. (2012). *Sampling* (3. ed.). *Wiley series in probability and statistics*. Wiley. https://doi.org/10.1002/9781118162934

Valliant, R., Dever, J. A., & Kreuter, F. (Eds.). (2018a). *Practical Tools for Designing and Weighting Survey Samples*. Springer, Cham.

Valliant, R., Dever, J. A., & Kreuter, F. (2018b). Variance Estimation. In R. Valliant, J. A. Dever, & F. Kreuter (Eds.), *Practical Tools for Designing and Weighting Survey Samples* (pp. 421–480). Springer, Cham. https://doi.org/10.1007/978-3-319-93632-1_15

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research*. University of Chicago Press. http://gbv.eblib.com/patron/FullRecord.aspx?p=557591

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

Wolter, K. M. (2007). *Introduction to Variance Estimation* (2nd ed.). *Springer series in statistics*. Springer New York.

Chapter 4: Sampling weights in multilevel modelling: an investigation using PISA sampling structures

Chapter 4 evaluates the best application of sampling weights in hierarchical models. Based on theoretical foundations and practical developments, we optimize the application of hierarchical model weighting in LSAs through simulative evaluation. We examine nine different weighting designs. The selected scenarios are based on framing approaches to explain required weighting in hierarchical modelling, settings promoted in the literature and theoretical, new devised considerations for future implementations. We consider different estimation, optimization, acceleration methods, and approaches to using sampling weights. Three population scenarios have been simulated using the statistical program R (R Core Team, 2018). The analyses have been performed with two software packages for hierarchical modelling of LSA data: Mplus (Muthén & Muthén, 2017) and SAS (SAS Institute Inc., 2018).

Contributing article:

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: an investigation using PISA sampling structures. *Large-scale Assess Educ 9*, 6. https://doi.org/10.1186/s40536-021-00099-0

Author contributions:

Mang devised the theoretical considerations, the research questions, derived and implemented the simulation study, conducted the data analysis, and wrote the first draft of the manuscript. All authors contributed to the interpretation of the results and the writing and revision of the manuscript.

Mang *et al. Large-scale Assess Educ* (2021) 9:6
https://doi.org/10.1186/s40536-021-00099-0

Large-scale Assessments in Education

METHODOLOGY Open Access

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

Julia Mang[1*], Helmut Küchenhoff[2], Sabine Meinck[3] and Manfred Prenzel[4]

*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 $\widehat{\theta}_{PML}$ is therefore design consistent for the finite population maximum likelihood estimator $\widehat{\theta}$, which, in turn, is model-consistent for the superpopulation estimator of θ. Therefore $\widehat{\theta}_{PML}$ 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 derivates 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 technically 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

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

Here, K is the negative second derivative of the logarithmic pseudo likelihood evaluated at $\widehat{\boldsymbol{\theta}}$. 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: Cluste*r. 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 propability 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

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

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 We ights
			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
			House weights
			Clustersum
	Model 3	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 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	$w_i{}^*$	$w_j{}^*$
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 $\widehat{\beta_0}$	CR $\widehat{\beta_1}$	CR $\widehat{\beta_2}$	CR $\widehat{\sigma_\varepsilon^2}$	CR $\widehat{\sigma_\tau^2}$
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 $\widehat{\beta_0}$	CR $\widehat{\beta_1}$	CR $\widehat{\beta_2}$	CR $\widehat{\sigma_\varepsilon^2}$	CR $\widehat{\sigma_\tau^2}$
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 $\widehat{\beta_0}$ 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* $\widehat{\beta_0}$. 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* $\widehat{\sigma_\varepsilon^2}$ 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* $\widehat{\sigma_\tau^2}$ of the

Table 6 Coverage rates of low variances simulated data

Software	Weighting approach	CR $\widehat{\beta_0}$	CR $\widehat{\beta_1}$	CR $\widehat{\beta_2}$	CR $\widehat{\sigma_\varepsilon^2}$	CR $\widehat{\sigma_\tau^2}$
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 $\widehat{\beta_0}$	CR $\widehat{\beta_1}$	CR $\widehat{\beta_2}$	CR $\widehat{\sigma_\varepsilon^2}$	CR $\widehat{\sigma_\tau^2}$
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* $\widehat{\sigma_\tau^2}$ 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 $\widehat{\sigma_\tau^2}$. 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 $\widehat{\beta_1}$ by the SAS software program with its procedure GLIMMIX although interquartile spaces include the true value and overlap with one another. However, this effect is offset by the addition of the average SES $\widehat{\beta_2}$ 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 $\widehat{\sigma_\varepsilon^2}$ decreases. This is caused by the increase in explained variance by adding the SES indicator. The same applies for the *Variance Between* the schools $\widehat{\sigma_\tau^2}$ 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 $\widehat{\beta_0}$	CR $\widehat{\beta_1}$	CR $\widehat{\beta_2}$	CR $\widehat{\sigma_\varepsilon^2}$	CR $\widehat{\sigma_\tau^2}$
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 $\widehat{\beta_0}$	CR $\widehat{\beta_1}$	CR $\widehat{\beta_2}$	CR $\widehat{\sigma_\varepsilon^2}$	CR $\widehat{\sigma_\tau^2}$
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 $\widehat{\beta_0}$ 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 ($\widehat{\beta_0}$) 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* $\widehat{\sigma_\varepsilon^2}$ 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 ($\widehat{\sigma_\varepsilon^2}$). A different picture as in Simulation Scenario 1 can be seen for the estimation of the *Variance Between* $\widehat{\sigma_\tau^2}$ in Simulation Scenario 2 (Fig. 2, Graph C, Graph G and Graph L). The *Variance Between* $\widehat{\sigma_\tau^2}$ 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

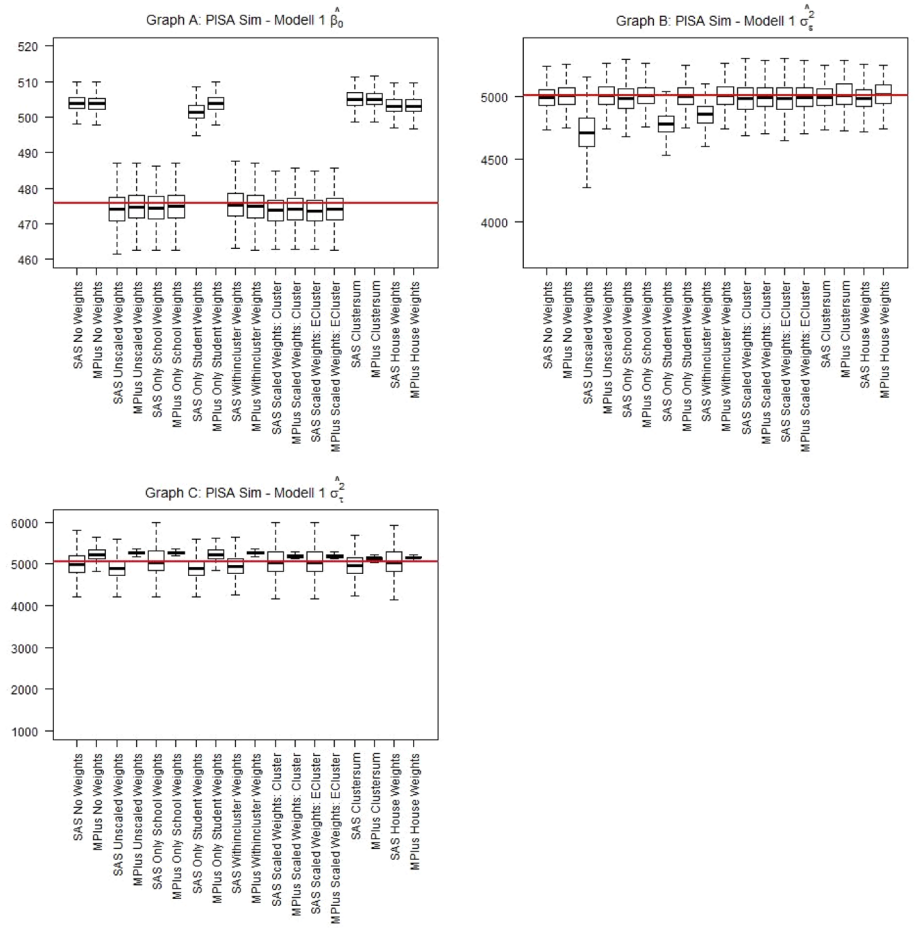

Fig. 1 Simulation outcomes for data mirroring the German PISA population. Model 1 is declared as $y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$, Model 2 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$ and Model 3 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$. The median, the 25% and 75% quartiles, minimum and maximum for each model estimator are presented in boxplots. No outliers are displayed. The true population values for each estimate are marked as red line. Simulation variation is displayed with the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages

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 $\widehat{\beta_1}$ 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 $\widehat{\beta_1}$ and school level $\widehat{\beta_2}$ 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

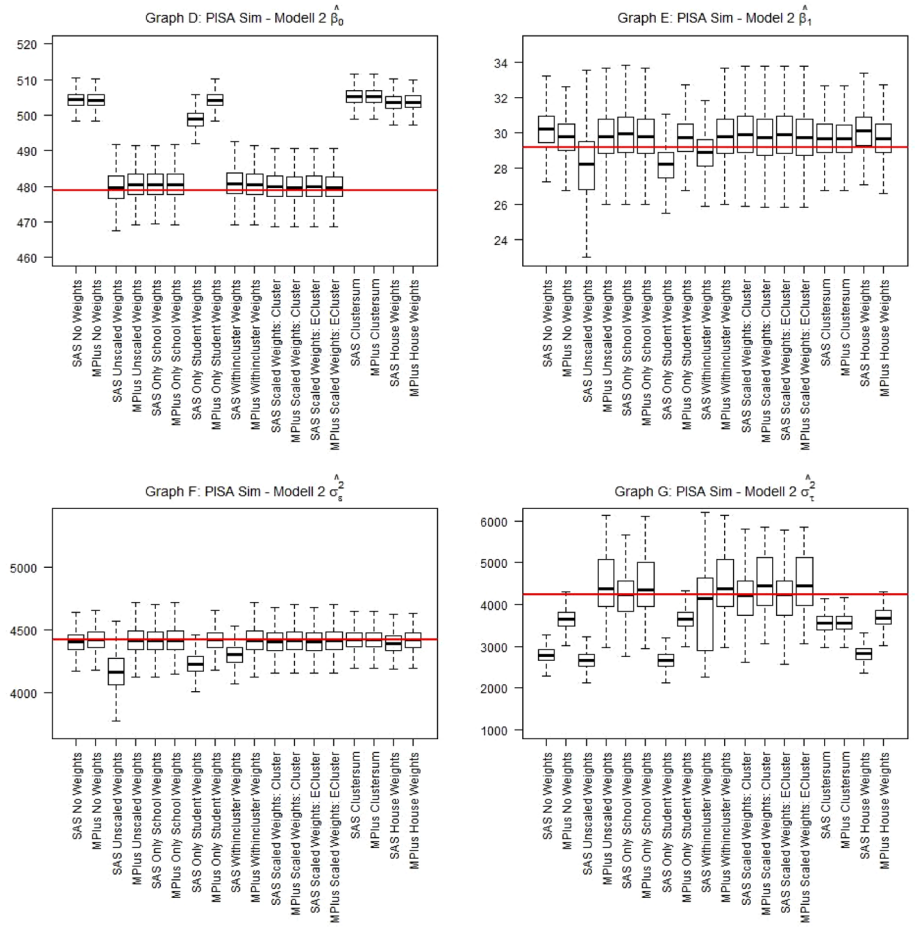

Fig. 1 continued

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 $\widehat{\sigma^2_\tau}$. 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* $\widehat{\sigma^2_\tau}$ 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 $\widehat{\beta_2}$ into the model, this difference disappears and estimators of the *Variance Between* $\widehat{\sigma^2_\tau}$ and the socio-economic background $\widehat{\beta_1}$ and $\widehat{\beta_2}$ 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.

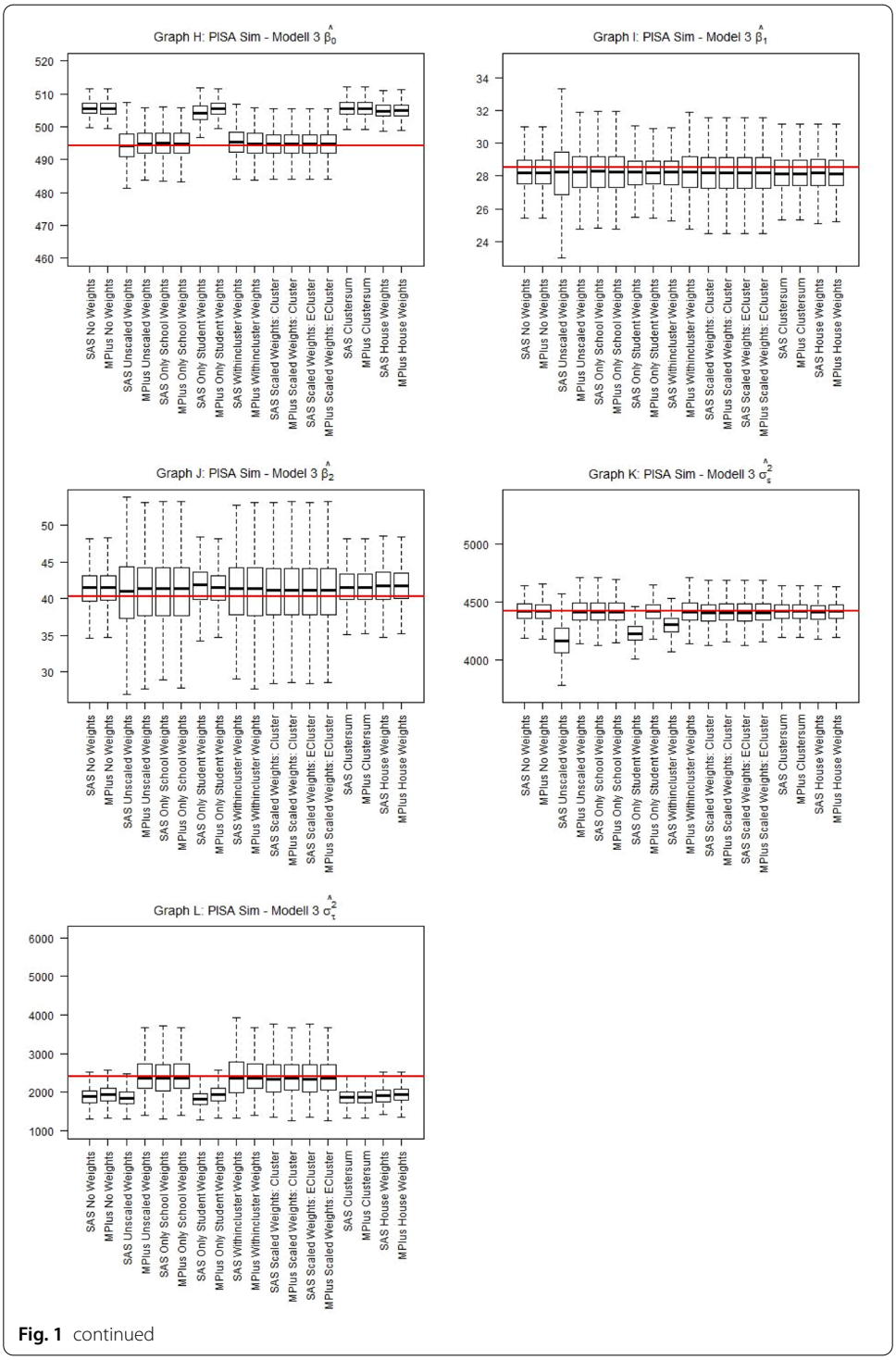

Fig. 1 continued

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 $\widehat{\beta_1}$). Although the confidence intervals are sometimes quite small, they are partly more

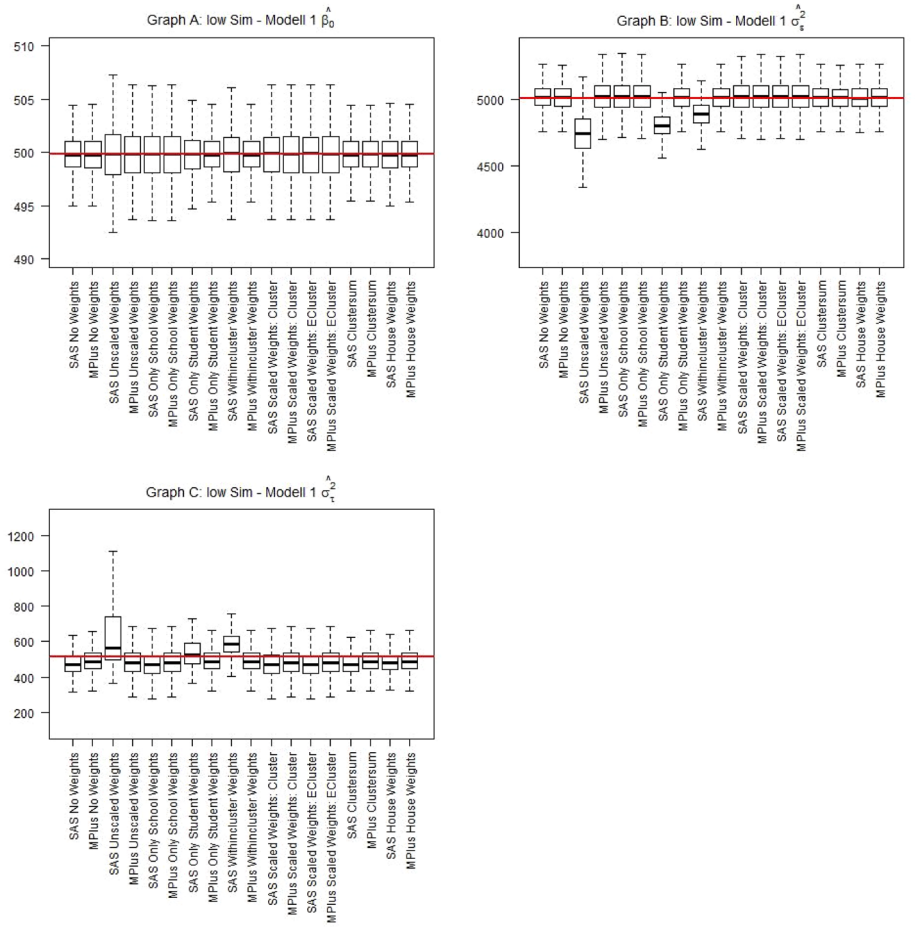

Fig. 2 Simulation outcomes for data reflecting low variances between schools. Model 1 is declared as $y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$, Model 2 *as* $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$ and Model 3 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$. The median, the 25% and 75% quartiles, minimum and maximum for each model estimator are presented in boxplots. No outliers are displayed. The true population values for each estimate are marked as red line. Simulation variation is displayed with the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages

biased (refer e.g., to Fig. 1, Graph L, or Table 5 $B\ \widehat{\sigma^2_\tau}$). 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

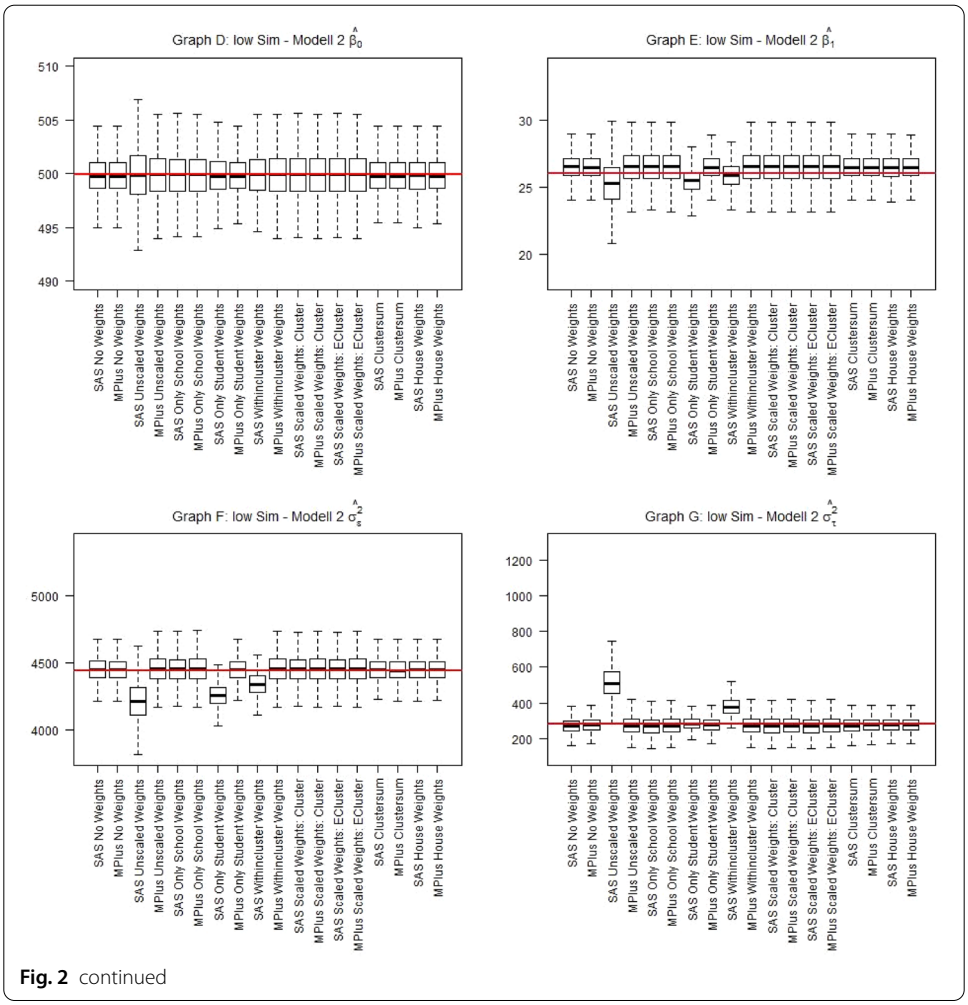

Fig. 2 continued

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

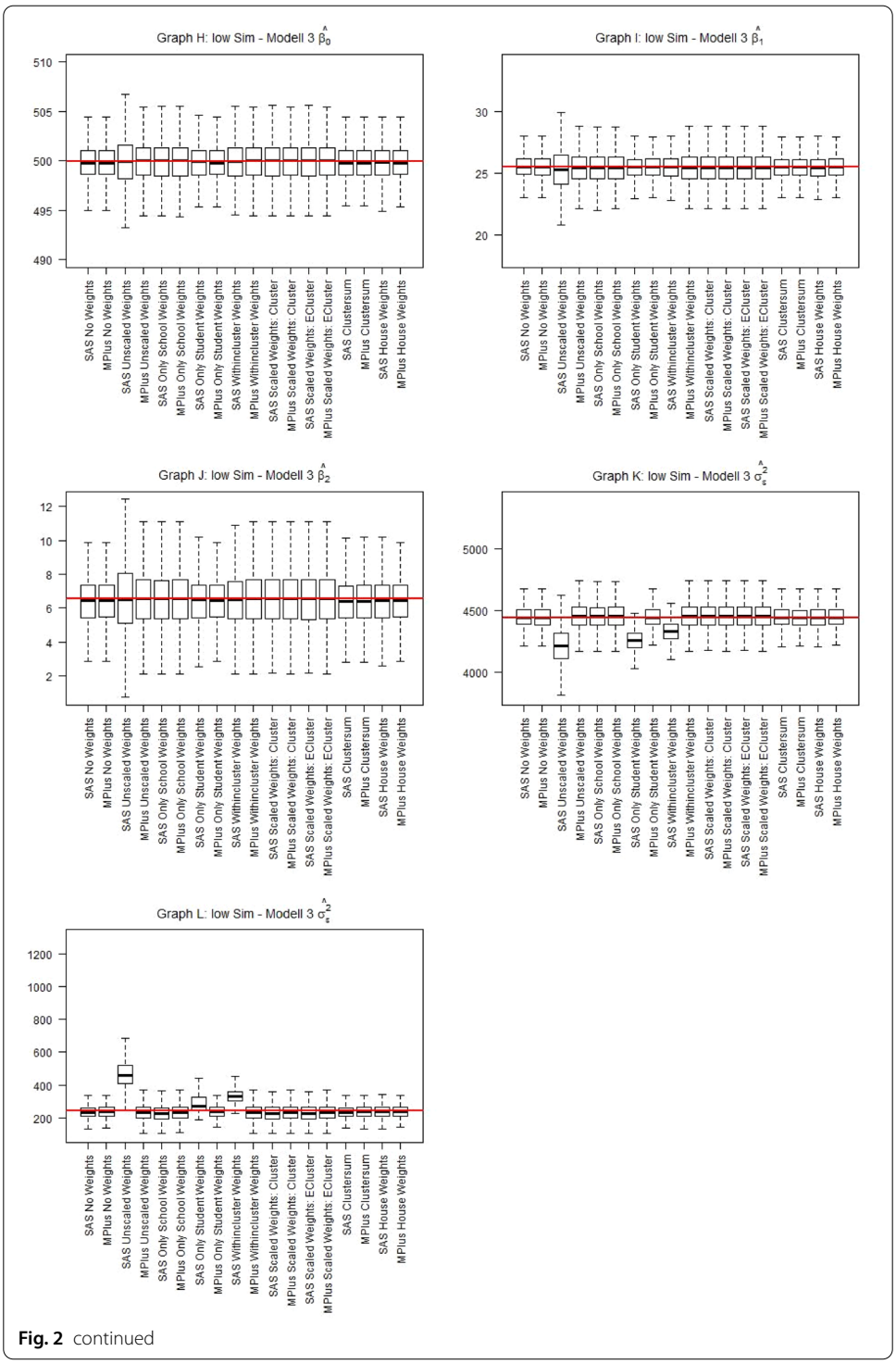

Fig. 2 continued

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

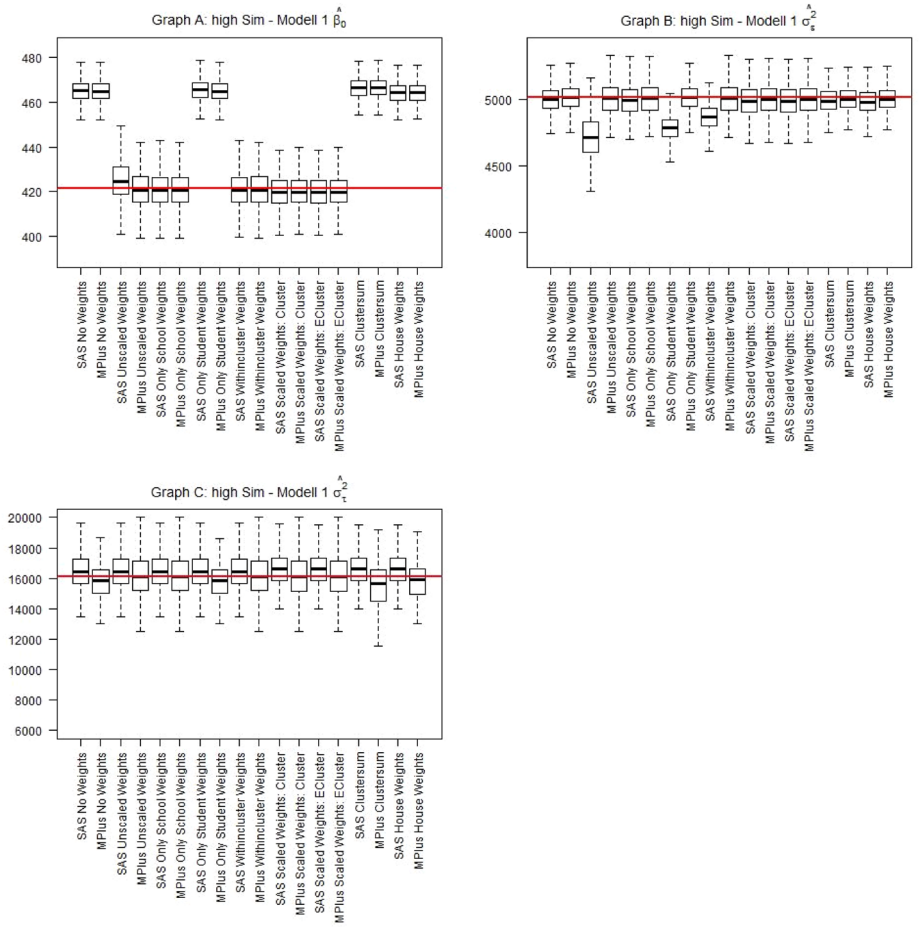

Fig. 3 Simulation outcomes for data reflecting high variances between schools. Model 1 is declared as $y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$, Model 2 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$ and Model 3 as $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$. The median, the 25% and 75% quartiles, minimum and maximum for each model estimator are presented in boxplots. No outliers are displayed. The true population values for each estimate are marked as red line. Simulation variation is displayed with the different weighting approaches combined with different estimation algorithms implemented in the two examined software packages

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

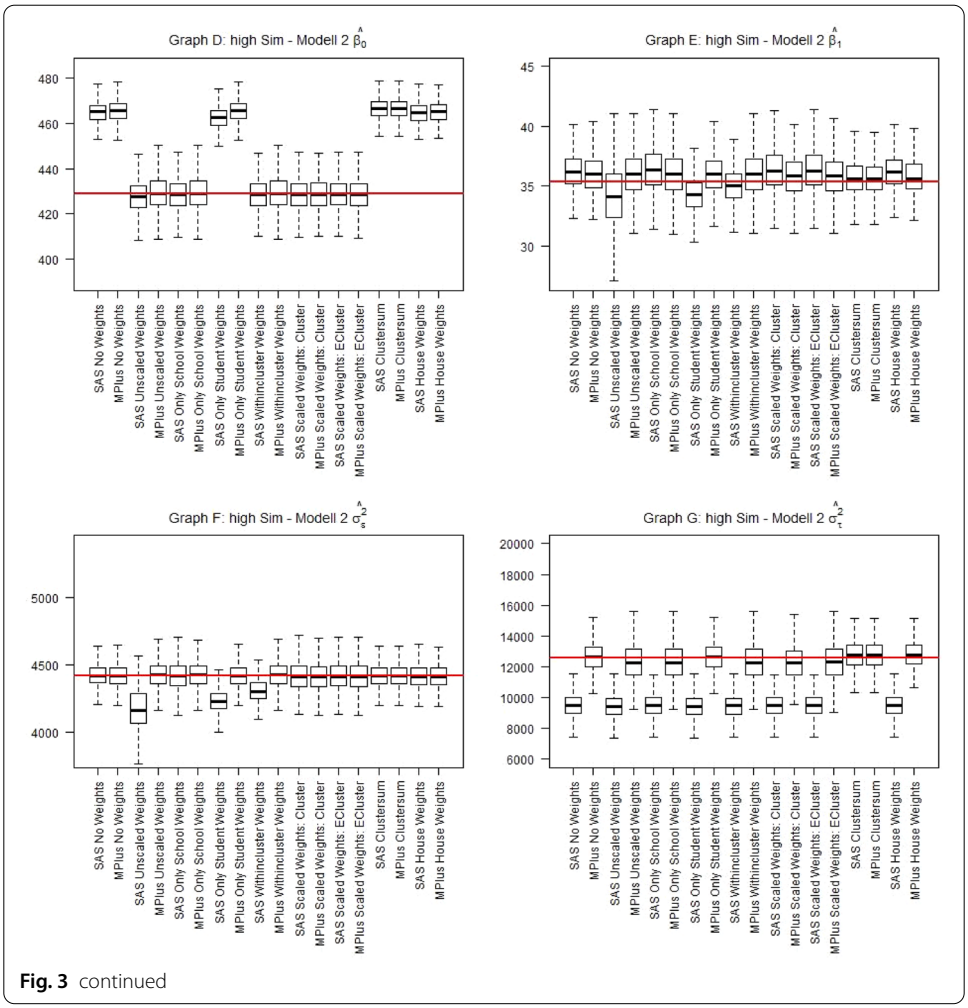

Fig. 3 continued

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.

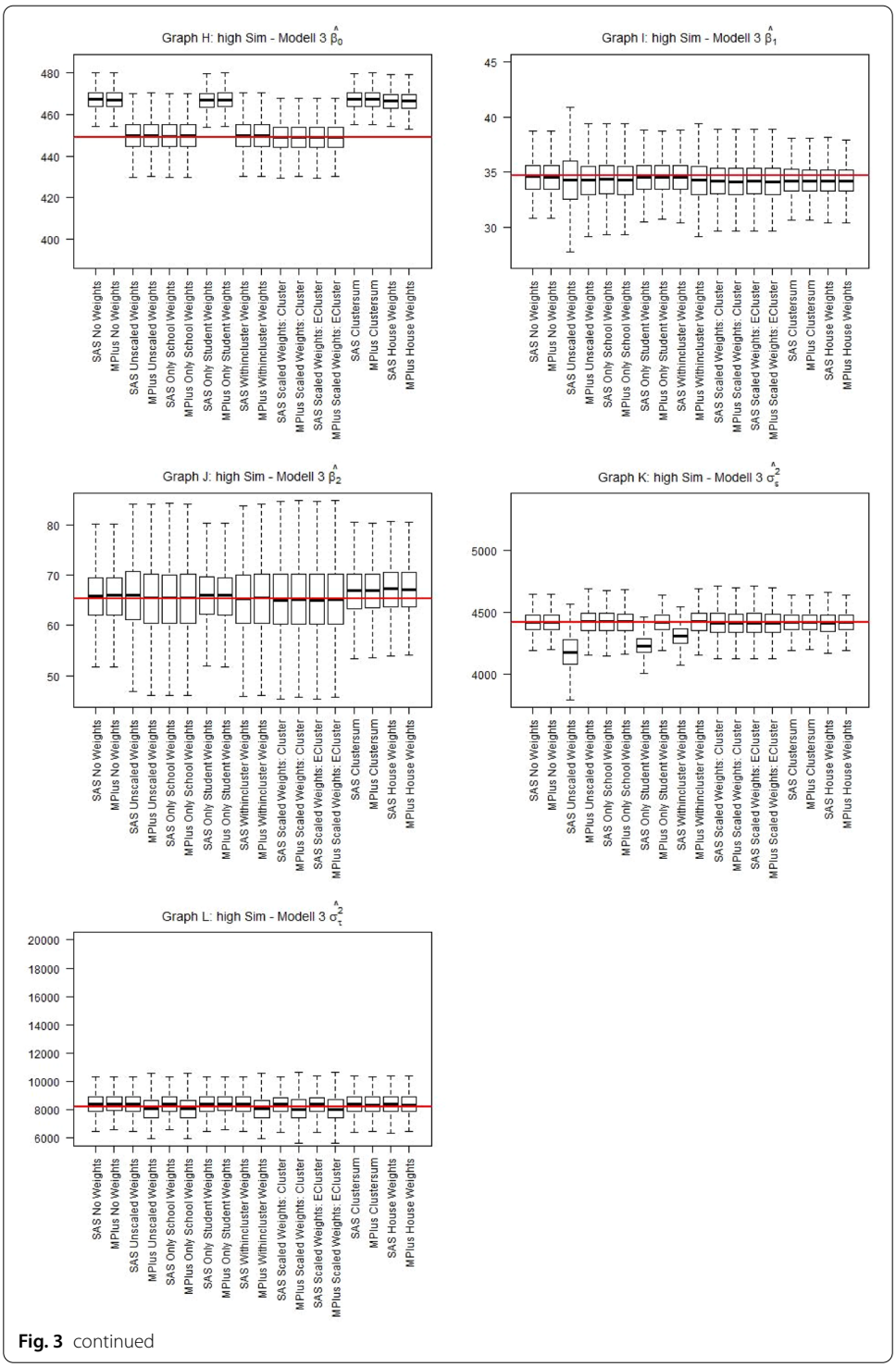

Fig. 3 continued

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

	$\widehat{\beta_0}$	SE $\widehat{\beta_0}$	$\widehat{\sigma_\varepsilon^2}$	SE $\widehat{\sigma_\varepsilon^2}$	$\widehat{\sigma_\tau^2}$	SE $\widehat{\sigma_\tau^2}$
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 $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \tau_i + \varepsilon_{ij}$

	$\widehat{\beta_0}$	SE $\widehat{\beta_0}$	$\widehat{\beta_1}$	SE $\widehat{\beta_1}$	$\widehat{\sigma_\varepsilon^2}$	SE $\widehat{\sigma_\varepsilon^2}$	$\widehat{\sigma_\tau^2}$	SE $\widehat{\sigma_\tau^2}$
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 $y_{ij} = \beta_0 + \beta_1 * x_{ij} + \beta_2 * x_i + \tau_i + \varepsilon_{ij}$

	$\widehat{\beta}_0$	SE $\widehat{\beta}_0$	$\widehat{\beta}_1$	SE $\widehat{\beta}_1$	$\widehat{\beta}_2$	SE $\widehat{\beta}_2$	$\widehat{\sigma^2_\varepsilon}$	SE $\widehat{\sigma^2_\varepsilon}$	$\widehat{\sigma^2_\tau}$	SE $\widehat{\sigma^2_\tau}$
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 (Davier 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 $\widehat{\beta}_0$ 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.
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 Survery 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.

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

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011 in der Fassung der 1. Änderungssatzung vom 6. Juni 2012 und der 2. Änderungsfassung vom 29. September 2016, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

_____

Julia Mang

München, 02.08.2023