Mang *et al.*
Large-scale Assessments in Education (2024) 12:15
https://doi.org/10.1186/s40536-024-00203-0

Large-scale Assessments in Education

METHODOLOGY Open Access

Evaluating German PISA stratification designs: a simulation study

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

*Correspondence:
Julia.Mang@goethe.de

[1] Goethe-Institut e.V., Oskar-Von-Miller-Ring 18, 80333 Munich, Germany
[2] Ludwig-Maximilians-Universität München, Institut Für Statistik, Ludwigstr. 33, 80539 Munich, Germany
[3] International Association for the Evaluation of Educational Achievement (IEA), Überseering 27, 22297 Hamburg, Germany

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 15-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 15-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 15-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 five 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 15-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 15-year-old students in Germany (Brown, 2010; Kish, 1965; Levy & Lemeshow, 2013; Thompson, 2012). To make correct inferences about the population of 15-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 15-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 15-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 (7th to 10th grade, born in the year of definition), the number of 7th to 10th 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

[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.

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 PISA, the preferred MOS is the expected number of 15-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. Please note that $nMOS_i$ can be greater than or equal to the sum of MOS_j in exceptional cases. These schools are then removed from the list and the probabilities for selection are re-estimated. As the simulation in this paper is based on an existing sample, such schools are not part of the simulation and therefore do not need to be taken into account. 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

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

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 subsamples 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

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

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", 2000 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 15-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

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	

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 15-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 15-year-old students, so that no testing could occur. Two other schools were excluded during the assessment (Mang et al., 2019). Table 10 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.

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

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	

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.

[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.

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

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, 2000 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 15-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 2000 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 15-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 a weight of 200 represents 200 students in the population.. 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 2000 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 2000

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

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

Method: linear regression (unadjusted R^2)

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

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

Method: linear regression (unadjusted R^2)

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 (Tables 5 and 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).

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

[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.

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 12 and 13 in 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 (2000 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 Fig. 1 and Fig. 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 2000 sample replicates. Column 4 displays the "true" SE for each variant, calculated as the standard deviation (SD) of the average student achievement over the 2000 sample replicates, i.e., the SD of the sampling distribution. Finally, we present

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

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

SE sampling error, *CR* coverage rate, *BRR* balanced repeated replication, *RMSE* root mean squared error

in column 5 Root Mean Squared Error (RMSE). A low RMSE value means that the estimator's bias and variance are small.

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. These minor deviations can be explained by the simulation of the population itself. In particular, if there are only a few students in a school or few schools in a stratum, deviations in mean estimations can occur.

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 (2000 samples). The red horizontal line represents the

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

SE sampling error, *CR* coverage rate, *BRR* balanced repeated replication, *RMSE* root mean squared error

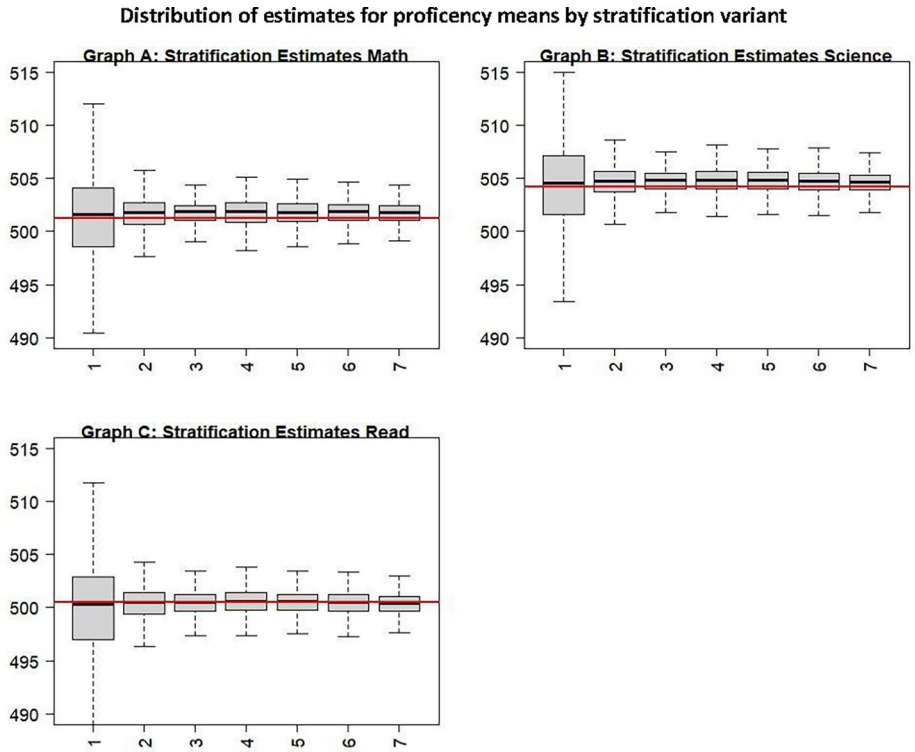

Fig. 1 Distribution of estimates for proficency means by stratification variant. Please refer to Table 4 for the description of the stratification variants

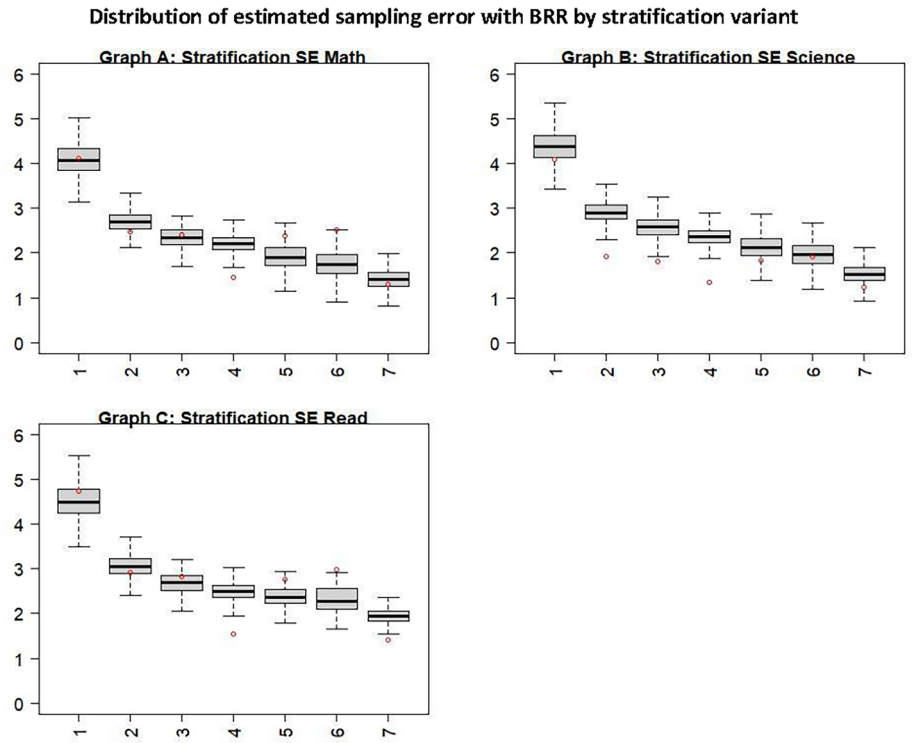

Fig. 2 Distribution of estimated sampling error with BRR by stratification variant. Please refer to Table 4 for the description of the stratification variants

true population value. Looking closely at Fig. 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.

This research focuses on sampling precision, which is presented in columns 3 and 4 of Tables 7, 8, 9, augmented by a graphical display (Fig. 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, 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, 8, 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 Fig. 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, 8, 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 2000 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 reperformed all analyses with a random permutation for the applied implicit stratification variables to address this hypothesis. Unlike the implicit stratification in stratification

[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 Tables 7,8,9 are due to the simulation design.

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 11 with 100 replicates in 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 Appendix; for the other domains, the outcomes are comparable.

Related to this, note that the coverage rates presented in columns 2 of Tables 7, 8, 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. Nevertheless, it must be kept in mind that a smaller sample size can also result in a smaller number of possible subgroups to analyze. 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.

Appendix

See Tables 10, 11, 12, 13.

Table 10 Comparison of the simulated school population and the true school population of PISA 2018 (Frame)

	Simulated school population	PISA 2018 school population (Frame)
N	13,046	13,855
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

Due to data protection reasons, strata were pseudonymized

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

* No school type 6 has been sampled in PISA 2018

Table 11 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[a]	–	–	–	–	–
6	0.40	0.97	2.38	2.99	3.00
7	0.20	0.97	2.15	2.70	2.69

SE sampling error, *CR* coverage rate, *BRR* balanced repeated replication, *RMSE* root mean squared error

[a] No implicit stratification for Stratification 5 variant

Table 12 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

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

Method: linear regression

Table 13 Explained variances in student proficiency (math, science, and reading) by explicit and implicit stratification for each stratification variant for real PISA 2018 sample data

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

Method: linear regression

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

Acknowledgements
Not applicable.

Author 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

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors gave permission for publication.

Competing interests
The authors declare that they have no competing interests.

Received: 10 October 2022 Accepted: 29 April 2024
Published online: 08 May 2024

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). *Sampling designs of the National Educational Panel Study: Challenges and solutions* (Vol. 14). VS Verlag fur Sozialwissenschaften. 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 P. L. Peterson, 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* (3rd ed.). A Wiley publication in applied statistics.
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, 6*, 223–239.
Kauermann, G., & Küchenhoff, H. (2011). *Stichproben: Methoden und praktische Umsetzung in R*. Springer. 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.
Lavrakas, P. (2008). Taylor series linearization (TSL). In P. Lavrakas (Ed.), *Encyclopedia of survey research methods.* (Vol. 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, 9*(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*. 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* (2nd ed.). 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*. pp 4257–4263.
Rizzo, L., & Rust, K. F. (2011). Finite Population Correction (FPC) for NAEP Variance Estimation. *JSM Proceedings: Survey Research Method Section*. pp 2501–2515.
RStudio Team. (2020). *RStudio: Integrated Development Environment for R (Version 1.4.1717) [Computer software]*. RStudio, Inc.
Rubin, D. B. (1993). Discussion statistical Disclosure Limitation. *Journal of Official Statistics, 9*(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., von Davier, M., & Rutkowski, D. (Eds.). (2013). *Handbook of international large-scale assessment*. 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. 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* (pp. 1–5). Wiley StatsRef: Statistical Reference Online.
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*. 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*. https://doi.org/10.18637/jss.v079.i10
Thompson, S. K. (2012). *Sampling. Wiley series in probability and statistics* (3rd ed.). 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.
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. https://doi.org/10.1007/978-3-319-93632-1_15
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*.
von Davier, M., & 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*. Chapman and Hall/CRC.
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. Springer series in statistics* (2nd ed.). Springer.

Publisher's Note