An Examination of the Performance of Variance Estimators in International Large-Scale Assessments

Final Report

2025.07.07

Umut Atasever, Sabine Meinck, Diego Cortes[1]

[1] All authors are affiliated with IEA, Hamburg, Germany.

Table of Contents

An Examination of the Performance of Variance Estimators in International Large-Scale Assessments

1 Abstract

The primary objective of this study is to examine the performance of the most widely used sampling variance estimators in international large-scale assessments (ILSAs): *the Balanced Repeated Replication (BRR)* method and *the Jackknife Repeated Replication (JK2)* method (both *Half* and *Full* variants). Additionally, we investigate the impact of *Fay* modification factors (*10%*, *30%*, and *50%*) on both *BRR* and *JK2* methods in estimating sampling variance. We also evaluate Bootstrapping as an alternative variance estimator for clustered sampling in ILSAs. Aiming for comprehensive results, we examine different conditions reflecting common scenarios in ILSAs. We vary school sample sizes, namely looking at samples of 30, 50, 100, and 150 schools. The study further explores the treatment of Primary Sampling Units (PSUs) in variance strata formation under school non-response scenarios, and the occurrence of odd numbers of PSUs in given explicit strata. A Monte Carlo simulation is conducted, mirroring a TIMSS Grade 4 student population with realistic distributions of achievement scores, standard deviation, intra-class correlation coefficient (ICC), and background characteristics. Probability samples are repeatedly drawn following a two-stage stratified cluster sampling design, where schools are PSUs and classes are secondary sampling units. One thousand samples are selected for each sample scenario. For each sample, the population parameter of interest is estimated, and for each sample scenario, its sampling variance is investigated. The "*true*" sampling error is approximated from the variability of estimates across the sample iterations as the standard deviation of the sampling distribution. The performance of each variance estimator is assessed by comparing the respective estimated sampling error, computed as the average of the square roots of the estimated sampling variances across all samples, to the (*approximated*) true sampling error. Results are summarized using *Relative Bias* to assess accuracy by quantifying over- or under-estimation and *Stability* to measure precision. Results indicate that *Bootstrapping, JK2* and *BRR without Fay* modification yield the most accurate sampling variance estimates for smooth statistics such as means, with *JK2* demonstrating the highest precision across conditions. For non-smooth statistics such as percentiles, *Bootstrapping* and *BRR* outperform *JK2* in terms of accuracy and precision. Notably, the *JK2 Half* and *Full* variants perform virtually identically in terms of variance estimation accuracy and precision. *Fay-modified JK2* does

not improve estimation precision, while *BRR with a Fay factor* improves performance for non-smooth statistics, particularly at 10% and 30%, compared to the traditional *50% Fay factor*. Under non-response conditions, specific methods currently in use for contemporary large-scale assessments consistently under- or overestimate sampling variance, suggesting that unadjusted variance strata introduce systematic bias. This study advances discussions on variance estimation in ILSAs, offering insights into the optimal application of resampling methods, including the trade-offs of *BRR*, *JK2* and *Bootstrapping* under varying statistical conditions and sampling designs.

Keywords: *large-scale assessments, complex sample designs, sampling variance estimators, replication methods, Monte Carlo simulation*

2 Introduction

International large-scale assessments (ILSAs) such as the Trends in International Mathematics and Science Study (TIMSS) [2] and the International Computer and Information Literacy Study (ICILS)[3] conducted by the International Association for the Evaluation of Educational Achievement [IEA] and the Programme for International Student Assessment (PISA)[4] coordinated by the Organization for Economic Co-operation and Development [OECD]) are essential for measuring national trends in educational achievement globally (Hernández-Torrano & Courtney, 2021). In these studies, probability sampling serves as the foundation for estimating parameters that characterize the distribution of student attributes and outcomes in a population of interest (Meinck & Vandenplas, 2021). The sampling process that generates the data is complex and can be summarized as a two-stage stratified cluster sampling design (Meinck & Vandenplas, 2021). In the first stage, schools, which serve as primary sampling units (PSUs) in ILSAs, are sampled using probability proportional to size (PPS) within a given explicit stratum (Snijders & Bosker, 2011). They are implicitly sorted by a measure of size (i.e., the number of targeted students), in which larger schools have a higher probability of selection in the PPS method (Rust, 2013). Schools may be explicitly or implicitly stratified based on specific attributes, such as school type (e.g., public vs. private), to improve sample efficiency by creating homogenous groups of schools and students or to control sample sizes for research purposes or precision (Lohr, 1999; Meinck, 2020). In the second stage, students or classrooms (groups of students receiving instruction at school) are typically selected using systematic random sampling (SRS), where each student within a school has an equal probability of selection (see for class sampling [Siegel & Foy, 2024] in TIMSS and student sampling [OECD, 2024]).

During data collection, the sampling units at both stages are also subject to non-response (i.e., schools and students [Atasever, Jerrim & Tieck, 2024]), which adds further complexity to the ILSA sampling process. After data collection, total final weights are assigned to the participating units, incorporating both base weights (derived from selection probabilities) and non-response adjustments (Siegel & Foy, 2024). Any statistical parameter of interest is estimated using the total final weight, with an associated uncertainty caused by

[2] https://www.iea.nl/studies/iea/timss
[3] https://www.iea.nl/studies/iea/icils
[4] https://www.oecd.org/en/about/programmes/pisa.html

the sampling procedures (Karakolidis et al., 2022). In ILSAs, the combination of complex sample design and weighting adjustments for nonresponding units, along with potential poststratification procedures, directly impacts the level of error in statistical estimates (see Kish, 1965 for complex survey sampling; Mang et al., 2024 for ILSA framework). Consequently, these factors influence the distributional properties of various test statistics used for inference about population parameters (Wu, 2010). Given this complexity, accurate and precise sampling variance estimation is crucial for ensuring reliable statistical inference (Rust & Rao, 1996). In ILSAs, the standard deviation of this distribution is central to the assessment of the precision of study findings, as it quantifies the uncertainty around the estimated parameter. Due to the multistage stratified cluster sampling design it is inappropriate to apply '*textbook*' variance estimators, relying on the assumption that the data was collected through simple random sampling (Rutkowski et al., 2010).

There are multiple approaches to address the challenges associated with variance estimation in complex sample surveys including linearization techniques (Deville, 1999). However, replication methods, also known as resampling techniques, are among the most widely used approaches in survey methodology (Valliant et al., 2018). To account for the uncertainty induced by complex survey design, replication methods such as *Balanced Repeated Replication (BRR)*, *paired Jackknife Repeated Replication (JK2)* and *Bootstrapping* have been developed (Bruch et al., 2011). These methods systematically manipulate total final weights to generate a given number of "*somewhat different subsamples and somewhat different sampling weights*" (Rust & Rao, 1996, p. 285), which in turn produce pseudo-estimates of the statistical parameter of interest. By analyzing the variability among these pseudo-estimates, replication methods estimate the sampling distribution of a parameter (Rust, 2013). The advantage of replication methods is their straightforward implementation, as they systematically incorporate replicate weights that reflect multi-stage probabilistic sampling and non-response calibration (Valliant et al., 2018). This makes them particularly well-suited for complex survey design, such as ILSAs.

Even though *Bootstrapping* is popular in survey and non-survey statistics (Valliant et al., 2018), *BRR* and *JK2* are the most frequently employed methodologies for estimating sampling variance in ILSAs (Meinck, 2020). Variance estimators vary in both methodology and their application across different ILSAs (see for example *BRR* in PISA [OECD, 2024], *JK2* in ICILS [Schulz, 2020] and in TIMSS [Fishbein et al., 2024]). Replication methods for variance estimation have been extensively analyzed both theoretically (see Efron, 1982; Rust & Rao, 1996; Bruch et al., 2011; Valliant et al., 2018) and through simulation-based

comparisons in survey methodology (see Kovar, 1985; Kovar et al., 1988; Judkins, 1990; Austin, 2024). Yet, previous literature has not thoroughly explored how different variance estimators perform within the ILSA studies framework. This apparent gap, in conjunction with the remarkable attention such studies receive in public and policy, underscores the importance of investigating this issue to enhance our understanding and improve ILSA survey methodology. By focusing on two widely used estimators—*BRR* and *JK2*—the central aim of our research is to provide insights into their performance across diverse conditions. For the *Jackknife* family, we also compare the *Half* and *Full* methods. Additionally, we evaluate *Bootstrapping* as an alternative variance estimator for clustered sampling in ILSAs. The study further investigates the impact of a *Fay* factor on *BRR* and *JK2* and its implications for sampling variance estimation. Moreover, we explore the treatment of PSUs when generating variance strata, particularly under school non-response scenarios and when an odd number of PSUs exist within a given explicit stratum.

To achieve our research objectives, we simulate a student population similar to those in ILSAs, mirroring the TIMSS Grade 4 achievement score distribution. Probability samples are repeatedly drawn from this population using a Monte Carlo simulation approach. The sampling plan in our simulation reflects a typical design used in ILSAs: a two-stage cluster design, with schools as the primary sampling units and classes as the secondary sampling units. For each generated sample, the population parameter of interest and its sampling variance are estimated by the methods under investigation. The "*true*" sampling error is approximated by the standard deviation of the population parameter estimates across all 1,000 sample iterations of each scenario.[5] Additionally, for each variance estimator, the average sampling error is computed as the average of square roots of the estimated sampling variances across all samples. The ratio of the average estimated sampling error to the "*true*" sampling error is referred to as the *Relative Bias*, which measures how accurately each method estimates sampling variance. Furthermore, the study evaluates the *Stability* of sampling error estimates to determine how consistently each method performs across samples.

This report begins with a theoretical framework detailing key concepts and terms related to estimating sampling variance in complex samples, such as variance estimators, replication methods, and differences in variance strata construction. We then review the

[5] The sampling error of a parameter is the standard deviation of the parameter distribution obtained across all possible samples given a specific sampling design. With 1,000 samples we can approximate the true sampling error with sufficient precision (Meinck & Vandenplas, 2021).

current state of research to identify gaps addressed by the study. The methods section describes the design of the simulation study, that is, parameters of the simulated population, the sampling design used to reflect specific sample scenarios, the variance estimators under investigation, and the evaluation criteria used to determine their accuracy and precision. Results are presented across various scenarios and sample sizes with a summary highlighting the key findings. The following discussion provides interpretations, recommendations, and considerations for practical applications. The report concludes with a summary of contributions, limitations, and directions for future research.

3 Theoretical Framework

3.1 Standard Errors in ILSAs

The reporting of estimated statistical parameters based on sample data is typically accompanied by the estimation of the sampling variance of the estimator used. The sampling error (SE), defined as the square root of the estimated sampling variance, quantifies the precision with which parameters are estimated and are fundamental for statistical inference (Fishbein et al., 2024).

In ILSAs, student achievement scores are represented by plausible values (PVs; Mislevy et al., 1992) because students do not complete all assessment items (see Huang, 2024; Rutkowski et al., 2010). The total standard error accounts for both measurement error (also often referred to as imputation error)—reflected by the standard deviation of the five plausible values—and sampling error (Rutkowski et al., 2010). For instance, in the context of ILSAs, in TIMSS 2023 (Von Davier et al., 2024) the reported mean achievement in the mathematics assessment for Germany is 524 (2.1), where the value in parentheses represents the standard error, reflecting uncertainty due to the sample and assessment design. The sampling variance is estimated separately for each plausible value, and the final sampling error is obtained by averaging these sampling variances and taking the square root of the result (Gonzalez, 2024). However, since the sampling distribution of one plausible value is in expectation the same as of another plausible value, the variation across PVs primarily comes from assessment design rather than sampling design. Given this distinction, in our study we abstract from any measurement problem in the estimation of parameters that arises from the assessment design and focus solely on the sampling distribution of estimated parameters. That is, we assume that achievement in the assessment is measured with certainty. This implies that the sampling error of an estimated parameter is equal to its standard error. For the remainder

of this report, we use the term sampling error to refer to this uncertainty. However, it is important to note that while sampling error is typically the dominant component of the standard error in ILSAs, readers should also consider imputation error when working with ILSA data (Cortés et al., 2025).

Standard errors are equally essential when estimating parameters that represent differences between sub-groups, such as gender comparisons. For example, in Germany, TIMSS 2023 (Von Davier et al., 2024) reports that boys achieve on average 530 (2.5) score points, while girls achieve on average 517 (2.5) score points. The observed 13 score points difference alone is not evidence of a statistical difference in mean scores between boys and girls in the population. To arrive to a meaning and informative conclusion about this difference, we compute a t-value statistic by dividing the difference in the estimated means by the standard error of the difference (Rasch et al., 2010):

$$t = \frac{13}{2.6} = 5 \tag{1}$$

Since a t-value of 5 exceeds the critical threshold of 1.96,[6] the difference in achievement between boys and girls is statistically significant at the 95% confidence level, given the degrees of freedom criteria is met (Johnson et al., 1992). This implies that if we were to repeatedly sample from the population, at least 95 out of 100 samples are expected to yield a difference falling into the range of the confidence interval that can be constructed for this difference. Given the importance of standard errors in statistical inference, it is crucial to understand how they are estimated and to assess the performance of different variance estimators in the context of ILSA survey methodology.

3.2 Variance Estimators

Variance estimation in complex sampling designs often relies on techniques such as *linearization* and *replication* methods. *Linearization*, also known as the *Taylor Series* method, is a common approach for estimating variance in complex sampling designs (Deville, 1999).

[6] In this example, we use the critical value of 1.96, which corresponds to a p-value of 0.05 in the z-distribution, instead of a t-statistic from a t-distribution with *df* degrees of freedom. We note that this is only an approximation, since the t-distribution converges to the z-distribution as the number of degrees of freedom increases. How good this approximation is, depends on the number of degrees of freedom we have for estimating the mean difference between both groups.

Linearization variance estimation involves estimating variance based on differences among weighted totals for PSUs, reflecting the variability between these clusters. The key idea is to approximate nonlinear estimators, such as ratios, odds ratios, or regression coefficients, using a *linearization* function (Demnati & Rao, 2010). However, standard *linearization* methods are not effective for estimating the variance of quantiles, such as the median (50^{th} percentile) or the first (25^{th} percentile) and third quartiles (75^{th} percentile) (Valliant et al., 2018). While specialized *linearization* techniques have been developed for this purpose, they are infrequently used in ILSAs and are therefore not included in this study (Rust, 2013).

Among replication-based methods, *BRR* and *JK2* methods are the most commonly used variance estimators in ILSAs (Meinck, 2020). While *Bootstrapping* is not widely applied in ILSAs, it is a well-established method in other fields, and it performs well for "*any and everything*" (Valliant et al., 2018, p. 456), compared to the *BRR* and *JK2*. Re-sampling methods are both built on the principles of resampling by systematically manipulating the estimation weights. These replication methods create subsamples, known as sampling variance replicates, by systematically selecting subsets of PSUs (Kovar, 1985). When a particular PSU is included in a replicate, all sampled units within that PSU are retained as part of the replicate (Rust & Rao, 1996). These replicates serve as the basis for estimating sampling variances by resampling the data multiple times (Judkins, 1990).

3.3 Variance Strata

In ILSAs, schools are first grouped into explicit strata, treating them as independent populations, during the school sampling. Within each explicit stratum, schools are listed and sorted by predetermined categories, a process known as implicit stratification (Siegel & Foy, 2024), and further sorted by size within those categories. Schools are then selected to participate in the study within each explicit stratum, each school having a probability of selection that is proportional to its size (i.e., PPS sampling). That is, in the context of ILSAs, schools typically serve as PSUs, as they represent the first sampling level. Based on the systematic sampling applied to this sort order, PSUs are paired into variance strata for variance estimation (Meinck & Vandenplas, 2021). As adjacent sampled schools are more likely to share similar traits, pairing these schools into variance strata is a logical approach that reflects these similarities (Gonzalez & Foy, 1996).

One goal of stratification is to make estimation more efficient, i.e., reducing sampling error. This can be achieved by using stratification variables that generate homogeneous

groups of students or schools. For example, students within schools with similar characteristics—such as private schools—often exhibit greater similarities in outcomes compared to those in different school types (Foy, 2005). Each explicit stratum is treated as a distinct population, from which a representative sample of schools is drawn (Rust, 2013). Variance strata are constructed as pseudo-strata for approximating the sampling variance within a given explicit stratum. This approach reflects the contributions of both between- and within-PSU variability to the estimation of sampling variance (Gonzalez & Foy, 1996).

To illustrate the process of forming variance strata, let S_i represent a sampled school, where i denotes the sorting order within a given explicit stratum. Schools have already been sorted by size (measured as the number of students, referred to as the Measure of Size, MOS) and other stratification variables and sampled based on PPS. With the first six schools shown as follows:

$$S_1 = 150, \quad S_2 = 135, \quad S_3 = 110, \quad S_4 = 90 \quad S_5 = 50, \quad S_6 = 42 \qquad (2)$$

where the values reflect the MOS of the four example schools. The variance strata are then constructed as follows:

$$VS_1 = \{\text{Schools } S_1 \ S_2\}, VS_2 = \{\text{Schools } S_3 \ S_4\}, VS_3 = \{\text{Schools } S_5 \ S_6\} \qquad (3)$$

Schools S_1 S_2 are paired together as they share similar characteristics, particularly in terms of size (MOS respectively 150 and 135) and other observable attributes. Similarly, Schools S_3 S_4, are paired (MOS respectively 110 and 90), as they are more comparable to each other than to schools in Variance Stratum 1, and so forth.

Using these variance strata, a set of so-called "*replicate weights*" is generated, and the final weights for units in each replicate are manipulated based on the specific replication method employed (Rust & Rao, 1996). A full population estimate is computed for each replicate weight, and its sampling variance is estimated by aggregating the variation across these replicate estimates (Choudhry & Valliant, 2004). In other words, the variation across these replicate estimates relative to the mean estimate serves as an approximation of the sampling variance (Judkins, 1990). The sampling error SE_ε is computed as the square root of the sampling variance, as represented in Equation 4 (Gonzalez, 2024):

$$SE_\varepsilon = \sqrt{k * \sum_{r=1}^{R} (\varepsilon_r - \varepsilon_0)^2} \qquad (4)$$

Let ε_r represent the statistic of interest estimated using replicate weight r, and let ε_0

denote the same statistic estimated using the total sampling weight. Define R as the total number of replicate weights, and k as a scaling factor conditional on the specific replication method employed in the ILSA. We explain the scaling factor k in the following section, *Replication Methods*.

A common assumption in complex survey sampling is that variance strata can be leveraged to estimate sampling variance, treating each variance stratum as a separate sub-population for variance estimation (Rust & Rao, 1996). This approach effectively captures the hierarchical and clustered nature of the data. For example, a "*rule of thumb,*" suggested by (Valliant et al., 2010) is to equate the degrees of freedom with the number of variance strata, provided the assumptions underlying the design are met. While replication methods vary in implementation, they operate on the same fundamental principle: constructing variance strata within explicit strata for each country, although original variance strata prior to data collection could be kept, or variance strata reformed due to non-response from schools, as will be further detailed in the following sections (Fishbein et al., 2024; OECD, 2024). Hence, replication methods differ primarily in how replicate weights are computed.

3.4 Replication Methods

3.4.1 Jackknife Repeated Replication (JK2)

The *Jackknife Repeated Replication* (*JRR* or the *Jackknife*) method has been widely used in survey methodology and holds historical significance as the first established replication-based variance estimation method. Originally developed to reduce estimation bias in serial correlation coefficients by Quenouille (1949), Quenouille (1956) extended the technique to an infinite population framework (Bruch et al., 2011, p. 22). Subsequently, Tukey (1958) further refined the concept by proposing that subsample estimators generated through the *Jackknife* approach could serve as reliable variance estimators (Bruch et al., 2011, p. 22). The *Jackknife* method includes several variations, tailored to different sampling designs. The *JK1* method, also referred to as the "delete-one" approach, is used in single-stage sampling, where each replicate is created by systematically omitting one unit at a time and adjusting the weights of the remaining units to represent the *Full* population (Valliant et al., 2018). Given a sample size of n, the *Jackknife* variance estimator is constructed using $n - 1$ replicates (Efron, 1982). For multi-stage sampling designs, such as those used in ILSAs, the *paired JK2* method is employed (Wolter, 2007). *JK2* is a stratified version of *JK1*. In this context, "*deleting a unit*" refers to removing an entire PSU, meaning all sampled

units within that PSU are excluded from a given replicate estimate (Valliant et al., 2018, p.445), while doubling the weight of the paired unit in this variance stratum. The *JK2* method, which is particularly well-suited for clustered samples, forms the basis of variance estimation in IEA studies, including TIMSS and PIRLS (Meinck, 2020). Given that *JK2* is better suited to the stratified sample designs used in ILSAs than *JK1*, our focus is exclusively on *JK2*. Within ILSAs, the *JK2* approach estimates the sampling variance as described below.

Let the final sampling weight be denoted as w, and the replicate weight as rw. For a given variance stratum containing two PSUs, the replicate weights are calculated based on the *Jackknife Replicate Indicator (JKREP)*, which takes values of 1 (for doubling the weight) or 0 for zeroing the weight (Fishbein et al., 2024):

$$rw_r = \begin{cases} 2 \cdot w_1, if\ JKREP\ =\ 1 \\ 0 \cdot w_2, if\ JKREP\ =\ 0 \end{cases} \quad (5)$$

The *JK2* method hence systematically adjusts replicate weights within variance strata. In each variance stratum, one PSU is included with its weight doubled $2 \cdot w_1$, while the other PSU is omitted with its weight set to zero $0 \cdot w_2$. This process is sequentially applied to each variance stratum. For each stratum, the replicate weights are adjusted similarly—one PSU is doubled included, and the other is deleted—while weights of sample units in all other variance strata remain equal to the *Full* sampling weight $rw = w$. Once the process is repeated for all variance strata R replicate estimates are generated. The sum of squared differences between these replicate estimates and the full-sample estimate is used to estimate the sampling variance, as outlined in *Equation 4*.

In ILSAs, the *JK2* method is implemented in two primary variants: the *Half* and *Full* methods (Meinck, 2020). While both variants follow the same variance strata construction principles, they differ in the number of replicate weights generated per variance stratum. In the *JK2 Half* method, a single replicate weight is created for each variance stratum (Schulz, 2011; Schulz, 2015). By contrast, the *JK2 Full* method generates two replicate weights per variance stratum (Fishbein et al., 2024). In the *Half* method, a random indicator (commonly referred to as $JKREP$ and randomly coded as 1 or 0) determines whether a PSU's final sampling weight is doubled or set to zero-ed. The *Full* method follows the same procedure by dropping one PSU for each replicate but applies an additional replicate with reversed $JKREP$ values, in which each PSU is assigned a doubled weight at least once across the replicates (Fishbein et al., 2024). The number of replicates equals to number of variance strata in the

Half method $\mathrm{VS} = \mathrm{R}$, in the *Full* method, the number of replicates is twice the number of variance strata $\mathrm{VS} = \mathrm{R} * 2$.

Historically, the TIMSS Technical Report (1995) justified the choice of the JK2 method for its computational efficiency and near-unbiased variance estimates for means, totals, and percentages in complex samples (Gonzalez & Foy, 1996). As the flagship study of the IEA, TIMSS's choice of *JK2* and its sampling design set a precedent for subsequent IEA studies, leading to the adoption of *JK2* in the Progress in International Reading Literacy Study (PIRLS),[7] the International Civic and Citizenship Education Study (ICCS),[8] and ICILS. In early cycles of TIMSS and PIRLS, *JK2 Half* was used for estimating sampling variance (Gonzalez & Foy, 2000). This method forms one randomly selected subsample from each variance stratum. In practice, this translates into computing 75 replicate weights out of 75 variance strata. *JK2 Half* method is also utilized in studies like ICCS and ICILS, employing the same number of replicate weights for variance estimation (Schulz, 2011; Schulz, 2015). However, TIMSS significantly changed in 2015, adopting *JK2 Full* as the preferred method (LaRoche & Foy, 2016). This transition involved employing 150 subsamples and drawing two subsamples from each variance stratum. By doing so, TIMSS aimed to achieve more precise and reliable sampling variance estimates (LaRoche & Foy, 2016). Additionally, as PSU sample sizes have increased, PIRLS 2021 and TIMSS 2023 expanded their variance strata from 75 to 125, increasing replicate weights from 150 to 250. The scaling factor k for the sampling error SE_ε differs between these two methods. In the *Full* method, the total number of replicate weights R, is twice the number of variance strata, leading to a scaling factor of $k = \frac{1}{2}$. In the *Half* method, where R equals the number of variance strata, the scaling factor remains k = 1 (see **Table 1** for summary).

3.4.2 Balanced Repeated Replication (BRR)

Balanced Repeated Replication (BRR), also referred to as *balanced half-sampling*, was introduced by McCarthy (1969) as a variance estimation method specifically designed for sampling designs in which two PSUs are selected per variance stratum. This approach is particularly useful in area probability samples, where one of the primary objectives is to ensure geographic dispersion of PSUs (Valliant et al., 2018, p.452). In the original *BRR*

[7] https://www.iea.nl/studies/iea/pirls
[8] https://www.iea.nl/studies/iea/iccs

proposal, from two PSUs in a single variance stratum two replicates are derived, with the total number of variance strata denoted as VS, and the total number of replicates as $R = VS * 2$. However, McCarthy (1969) introduced an efficient computational method to reduce the number of required replicates. This method ensures that, for linear estimators, it estimates the "*same*" variance estimate that would be obtained by selecting all possible half-samples ($VS = R * 2$), but with significantly fewer replicates ($VS = R$ [Valliant et al., 2018, p.452]). This computational efficiency is achieved using an orthogonal Hadamard matrix of dimension HxH, where H is the smallest multiple of 4 that is greater than or equal to the number of variance strata (McCarthy, 1969). The Hadamard matrix provides a systematic method for structuring replicate weights.

The Hadamard matrix consists of $+1$ and -1 values, which dictate whether a PSU's weight is inflated or excluded in a given replicate. If a PSU is included, its weight is doubled, $f_u = 2$ and if a PSU is excluded, its weight is set to zero $f_d = 0$. For each variance stratum, half of the PSUs are systematically excluded (set to zero weights), while the other half have their weights inflated by a factor of two (Bruch et al., 2011). This process is similar to the Jackknife method, where PSUs are systematically removed, however, within a single variance stratum in each replicate estimate. The Hadamard matrix structure replicates, with each column representing a replicate and each row corresponding to a PSU within a variance stratum (Valliant et al., 2018, p.452). For example, in a 4 x 4 matrix, the first column includes all PSUs $+1$, while subsequent columns alternate PSU inclusion $(+1)$ and exclusion (-1).

$$H_4 = \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \quad (6)$$

In ILSAs, variance strata for *BRR* are constructed based on sampling attributes, and each PSU is assigned a replicate indicator, denoted as $BRREP$, which randomly takes values of 1 or 0. Replicate weights (rw) are then determined based on the combination of $BRREP$ and entries from a Hadamard matrix (H). If there are four variance strata VS = 4, which equals the dimension of the Hadamard matrix, then we generate four replicate weights R = 4. For a PSU in variance stratum vs and replicate r, the replicate weight $rw_{vs,r}$ is computed as:

$$rw_{vs,r} = \begin{cases} f_u \cdot w_{vs}, if\ BRREP = 1\ and\ H_{vs,r} = +1\ OR\ if\ BRREP = 0\ and\ H_{vs,r} = -1 \\ f_d \cdot w_{vs}, if\ BRREP = 0\ and\ H_{vs,r} = +1\ OR\ if\ BRREP = 1\ and\ H_{vs,r} = -1 \end{cases} \quad (7)$$

Where w_{vs} represents the final sampling weight for the PSU.

The traditional *BRR* estimates variance using half-sample replicates derived from the Hadamard matrix. The sampling error is computed using $Equation\ 4$, where the scaling factor is $k = \frac{1}{\mathrm{R}}$ and R equals the number of replicate weights (Judkins, 1990).

One limitation of the *standard BRR* is that half of the sample is excluded in each replicate estimate, which may cause instability in variance estimation, particularly for small subgroups or domains. If a domain exists in only a subset of the PSUs, there is a risk that all units within that domain could be excluded from certain replicates. An extension of the *BRR* method (*Fay*, 1984; Dippo et al., 1984; Judkins, 1990) proposed to multiply the estimation weights by a factor of f and $2 - f$ to solve the problem of missing replicates when approximating the sampling distribution of an estimator. This perturbation factor is controlled by the parameter f, where $0 \leq f < 1$. When $f = 0$, the method functions as the *standard BRR*. The perturbation factor of 50% *Fay* ($f = 0.5$) was proposed by *Fay* (1984) and Dippo et al (1984). If $f = 0.5$, the weights for PSUs in the half-sample are multiplied by 1.5, and while those excluded from the half-sample are multiplied by 0.5. This modification ensures that all PSUs contribute to variance estimation across all replicates.

Judkins (1990) conducted simulations to compare various *Fay* factors—specifically of 10% ($f = 0.1$), 30% ($f = 0.3$) and 50% ($f = 0.5$), and found that a *30% Fay* factor provided the most accurate and precise variance estimates. As proposed by *Fay* (1984), Dippo et al. (1984), and Judkins (1990), the squared differences formula ($Equation\ 4$) applies to *BRR with Fay* (also referred to as *Fay's Method*). However, in addition to the traditional *BRR* scaling factor, $k = \frac{1}{\mathrm{R}}$ where R equals to the number of replicate weights, an additional adjustment is introduced $\frac{1}{R*(1-f)^2}$. This additional scaling factor was suggested by Dippo et al. (1984) to ensure "*a reasonable estimate of the variance*" (Judkins, 1990, p. 225). For example, if $f = 0.5$ (*50% Fay factor*), the adjustment factor becomes $\frac{1}{R*(1-0.5)^2} = \frac{4}{\mathrm{R}}$. Without this correction, the mean squared error (MSE) of the variance estimate would be much smaller than expected.

In ILSAs, *BRR*'s variance strata are formed similarly to those in *JK2*, following the same sampling attributes, number of strata, and a replicate indicator variable, typically named *JKREP* or *BREP* (coded as 1 or 0) in the ILSA datasets. As we summarized above, the *BRR*'s

replicate weights are based on the Hadamard matrix. However, there are also differences between *BRR* and *JK2* in the formation of variance strata, particularly concerning odd number of PSUs in a given explicit stratum due to sample size or non-response. Section "*Variance strata differences*" below will focus on these differences of this paper.

In ILSAs where the *BRR* is used to estimate the sampling variance, a *50% Fay* factor is most commonly used. This approach is employed by PISA, the Teaching and Learning International Survey (TALIS)[9] conducted by OECD, and the Teacher Education and Development Study in Mathematics (TEDS-M),[10] organized by IEA. For example, PISA applies *Fay's Method* using 100 replicate weights, corresponding to 100 variance strata (OECD, 2024). TALIS follows the same approach (OECD, 2020). However, IEA's TEDS-M, generated only 32 replicate weights from 32 variance strata as the PSU sample size of the study was relatively small (Dumais & Meinck, 2013). The use of *BRR* in ILSAs is primarily due to its ability to provide stable estimates for non-smooth statistics (OECD, PISA 2009 Technical Report; OECD, TALIS 2018 Technical Report). In TIMSS 1995, where the *JK2* was commonly used, *BRR with 50% Fay* was also utilized for median, and percentile estimates due to its consistency for non-smooth statistics. However, subsequent cycles, such as TIMSS 1999, adopted only the JK2 method (Gonzalez & Foy, 2000). However, IEA also applied *BRR with 50% Fay* in TEDS-M (Dumais & Meinck, 2013), a study with relatively small samples, arguing that the method is better suited to this design.

Meanwhile, the *30% Fay* factor (f = 0.3) has been adopted in Programme for the International Assessment of Adult Competencies (PIAAC),[11] although its application varies across participating countries (OECD, 2012). PIAAC is a particularly notable case, as it employs multiple variance estimators, including *JK1*, *JK2*, *standard BRR*, and *BRR* with different *Fay* factors (*30%* and *50%*), depending on country-specific sampling considerations (OECD, 2012). We summarize in **Table 1** the variance estimators that are used in major ILSAs.

[9] https://www.oecd.org/en/about/programmes/talis.html
[10] https://www.iea.nl/studies/iea/teds-m
[11] https://www.oecd.org/en/about/programmes/piaac.html

Table 1 : Overview of Replication Methods for Selected ILSAs

ILSA	Replication Method	Fay Factor f	Scaling Factor k	Variance Strata VS	Replicate Weights R
TIMSS	JK2 Full	-	$\frac{1}{2}$	125	250
PIRLS	JK2 Full	-	$\frac{1}{2}$	125	250
PISA	FAY's BRR	0.5	$\frac{1}{R*(1-f)^2}$	100	100
ICCS	JK2 Half	-	1	75	75
ICILS	JK2 Half	-	1	75	75
TEDS-M	FAY's BRR	0.5	$\frac{1}{R*(1-f)^2}$	32	32
TALIS	FAY's BRR	0.5	$\frac{1}{R*(1-f)^2}$	100	100

3.4.3 Bootstrapping

The *Bootstrapping* (or *Bootstrap*) method, originally introduced by Efron (1982), is widely used in survey statistics due to its simplicity and adaptability. A notable adaptation for complex survey designs is the *Rao-Wu-Yue-Beaumont Bootstrapping* (Beaumont & Émond, 2022), which extends traditional *Bootstrapping* (Rao & Wu, 1988) to support multistage stratified designs and PPS sampling. This method generates replicate weights by systematically adjusting final sampling weights, and offers flexibility in the number of replicates. For instance, it can produce 200 replicate weights from 100 variance strata. Additionally, it explicitly accounts for selection probabilities for the sampling units when manipulating the replicate weights. While the *Bootstrapping* method is commonly employed in social and household surveys (Valliant et al., 2018), often alongside *Jackknife* (e.g., *JK2*) and *BRR* methods, it has not been widely adopted in ILSAs. However, its potential suitability for variance estimation in ILSAs warrants further investigation. In this study, we explore its application within ILSAs and compare its performance to *JK2* and *BRR* variance estimation methods. A detailed explanation of replicate weight construction and the implementation process is provided in the "*Methods*" section.

3.5 Differences in Building Variance Strata in ILSAs

3.5.1 Non-Response

A common assumption in the literature is that replication methods perform similarly under ideal conditions, such as large sample sizes and no complexities like non-response (Valliant et al., 2018). However, differences in variance strata formation arise when non-response occurs, particularly in ILSAs, impacting variance estimation. PSUs are subject to non-response in these studies, raising questions about how to handle participating versus non-participating PSUs.

3.5.1.1 TIMSS, PIRLS, and TALIS

In ILSAs such as TIMSS (Siegel & Foy, 2024), PIRLS (Foy & Almaskut, 2023), and TALIS (OECD, 2020), variance strata are constructed based solely on participating schools. PSUs (schools) are sorted by the same sampling attributes with which they were sorted prior to sampling (e.g., size, implicit stratification variables), and non-responding schools are excluded. The variance strata are defined as:

$$S_s = \{S_1, S_2, \dots, S_n\}, \quad \text{where } S_i \text{ is a participating school.}$$

Participating schools are paired sequentially to form variance strata. For example, if the responding schools are S_1, S_3, S_4, and S_6, they are paired as:

$$VS_1 = (S_1, S_3) \quad \text{and} \quad VS_2 = (S_4, S_6) \tag{8}$$

given they belong to the same explicit stratum and follow the sampling order.

3.5.1.2 PISA

In contrast, OECD's PISA (OECD, 2012; OECD, 2023) constructs variance strata at the time of sampling, before data collection. If a stratum originally contains two sampled schools, such as S_1 and S_2, and only S_1 responds, the remaining school is not re-paired. The variance stratum remains:

$$VS_1 = S_1, S_2 \tag{9}$$

In this case, replicate weights for VS_1 are generated only from the participating PSU, S_1. This approach avoids creating new pairs based on participation status and retains the original sampling structure.

The primary difference in these approaches lies in whether variance strata are redefined after data collection or remain fixed based on the original sampling design. This investigation examines whether retaining the original stratum structure, a method applied, e.g., in PISA, yields comparable outcomes to the re-pairing approach applied in TIMSS and TALIS.

3.5.2 Odd number of PSUs in the explicit stratum

When an explicit stratum contains an ***Odd Number of Schools***, *BRR* and *JK2* approaches differ in how they handle the unpaired PSUs (i.e., schools), after all PSUs are paired based on the sampling sort order. In *JK2*, the last unpaired school S_k is divided into two-quasi-PSUs to ensure that the variance stratum includes two PSUs (Fishbein et al., 2024). If S_k has one sampled class, students are randomly divided into two groups, forming quasi-schools Q_1 and Q_2, such that:

$$Q_1 \cup Q_2 = S_k, Q_1 \cap Q_2 = \emptyset \quad (10)$$

If S_k has multiple sampled classes, each class C_i is treated as a separate quasi-PSU, which allows the affected variance stratum to have two PSUs.

In *BRR*, the last unpaired school S_k is grouped with two preceding schools S_{k-1} and S_{k-2} to form a triplet within the variance stratum (Mang et al., 2024). Based on the Hadamard matrix and *BRREP* combinations, one school may receive an inflated adjustment factor, while the other two schools receive reduced factors (OECD, 2024). For instance:

$$f_k = 1.7071, f_{k-1} = f_{k-2} = 0.6464 \quad (11)$$

Alternatively, one school may receive a reduced adjustment factor, while the other two schools are inflated (OECD, 2024):

$$f_k = 0.2929, f_{k-1} = f_{k-2} = 1.3536 \quad (12)$$

These adjustment factors alternate across replicates based on the Hadamard matrix structure, ensuring that all schools contribute to the replicate weights.

By dividing the unpaired orphan school into quasi-schools in *JK2* or balancing triplet contributions in *BRR* through structured adjustment factors, both methods address odd numbers of schools in explicit strata.

3.6 Current State of Research

Survey methodologists have debated the relative performance of *JK2*, *BRR*, and *Bootstrapping*, often reaching differing conclusions despite well-established theoretical foundations (See Efron, 1982; Kovar, 1985; Kovar et al., 1988; Judkins, 1990; Rust & Rao, 1996; Bruch et al., 2011; Valliant et al., 2018; Austin, 2024). A commonly stated key advantage of these replication methods is that they are readily available for a wide set of statistical parameters and straightforward to implement using computational power, and their ability to account for selection probabilities and non-response adjustments by systematically manipulating final weights, particularly important in the ILSA context (Rust & Rao, 1996).

As highlighted by Valliant et al. (2018), replication methods, including *JK2*, *BRR*, and the *Bootstrapping*, rely on certain assumptions to provide unbiased and consistent variance estimates. These assumptions typically require sufficiently large sample sizes, no extreme outliers in the analysis variable, and the ability to compute necessary derivatives for nonlinear estimators (Valliant et al., 2018). While specific conditions may vary, the fundamental principle is that replication methods must align with the underlying sampling design and statistical properties of the data (Judkins, 1990). In other words, the way replicates are formed should accurately reflect the sampling methodology used to draw the sample from the population (i.e., stratification). Therefore, decisions regarding the construction of variance strata are crucial in understanding the performance of replication methods in ILSAs. This study examines this phenomenon to hypothesize potential sources of variation among different variance estimators (see section "*Methods*").

According to Valliant et al. (2018, p. 447), the *Jackknife* is theoretically equivalent to linearization in large samples for linear estimates (see Krewski & Rao, 1981 for theoretical conditions). This makes it particularly suitable for large-sample surveys such as ILSAs. Empirical findings confirm this. Judkins (1990), in a comprehensive simulation study, demonstrated that *JK2* has lower bias, greater *Stability*, and more precise confidence intervals compared to *BRR* (*with or without Fay*) in high-variance populations. Similarly, Kovar (1985) found that *JK2* tends to outperform *BRR*. However, theoretical discussions raise concerns about *JK2*'s performance for non-smooth statistics. Efron (1982) argued that *JK1* yields unreliable variance estimates for non-smooth statistics, including medians (50th percentile) and quartiles (25th and 75th percentiles). However, Kovar, Rao, and Wu (1988) empirically confirmed this, reporting that the *Jackknife* with stratification (*JK2*) performs worse than *BRR* for non-smooth statistics, in alignment with Efron's (1982) theoretical claims. Conversely,

adding *Fay's Method* to BRR consistently produced better variance estimates for non-smooth statistics compared to *JK2* and *BRR without Fay*. A noted limitation of *BRR*, however, is its reliance on the a priori calculation and storage of replicate weights due to the Hadamard matrix, whereas *Jackknife* weights can be generated systematically "*on the fly.*" Another limitation of the *BRR*, as noted in the sections before, is the number of missing replicate weights in small domains. However, this is solved by the *Fay's Method*. Judkins summarizes *Fay's 50%* suggestion as a "*compromise*" between *standard BRR* and *JK2*, but the *Fay's 30%* (which is a "*gentler*" perturbation of weights) provides more reliable estimates for quantile estimates. Judkins (1990) also found that *Fay's Method* is equivalent to the *standard BRR* in linear estimators. The application of the *Fay* factor in the *JK2* family remains unexplored in the literature. Compared to the original *BRR*, *JK2* does not face the issue of missing replicates to the same extent. However, the missing replicate challenge may still arise when sub-group analyses involve small sample sizes. For instance, the sampling variance of the mean estimator could be potentially omitted if a sub-group is confined within a pseudo-variance stratum. Introducing a *Fay* factor to *JK2* could address this limitation in sub-group analyses. Also, we expect that a *Fay* factor may enhance the performance of *JK2* for non-smooth statistics. However, this expectation warrants further investigation.

As highlighted by Valliant et al. (2018), the *Bootstrapping* extends beyond variance estimation to approximate the *Full* distribution of a statistic. This is accomplished by generating numerous subsamples and summarizing the resulting estimates, such as *BRR* and *JK2*. While the *Bootstrapping* shares general assumptions with other replication methods, specialized approaches, such as those described by Rao and Wu (1988) and Beaumont and Émond (2022), allow it to handle designs with multi-stage sampling, where the PSUs have varying selection probabilities, such as PPS. The *Bootstrapping*, particularly *the Rao-Wu-Yue-Beaumont* variant, is noted for its flexibility. However, these extensions are not widely available in standard software applicable to ILSA analysis. Currently, *Bootstrapping* can be implemented using R (R Core Team, 2022) with *svrep* (Schneider, 2023) and *survey* packages (Lumley, 2004), whereas *JK2*, *BRR*, and *Fay's BRR* with ILSA data can also be computed using tools such as *IEA's IDB Analyzer* (Sandoval-Hernandez & Carrasco, 2020; IEA, 2025), *RALSA* (Mirazchiyski, 2021), *EdSurvey* (Zhang et al., 2024), and *REPEST* (Keslair, 2020). Since *Bootstrapping* is widely used outside of ILSAs, this study explores whether it could be a viable alternative for variance estimation in ILSAs.

To date, research comparing the performance of different replication methods in ILSAs is relatively sparse. Practical proof, for example, based on a comprehensive simulation

study, is missing. Existing research has been largely theoretical on comparisons, primarily reported in technical documents such as the TIMSS Technical Report (1995) and PISA Technical Report (2000). Additionally, Rust (2013) presents a useful theoretical summary of replication methods used in ILSAs.

4 The Current Study: Research Questions

Considering the diverse variance estimation methods employed across major ILSAs, our research aims to conduct a comprehensive comparative analysis to assess the performance of sampling variance estimators across diverse conditions in the context of ILSAs. By examining widely used replication methods—*BRR*, *JK2* (*Full* and *Half* variants), and Bootstrapping—we investigate their effectiveness in estimating sampling variance for both smooth and non-smooth statistics, such as means, medians, quartiles, benchmark percentages, and correlations. Judkins (1990) has suggested that *a Fay factor of 30%* yields superior results for *BRR*, particularly for non-smooth statistics. Hence, we examine whether applying this modification enhances the precision of sampling variance estimation for such parameters. On the other hand, a *Fay* factor can take any value between 0 and 1. Further, scholars have argued that "*gentler*" perturbation of weights can lead to better *Stability* of the sampling variance estimator (see, for example, Judkins [1990, p.225]). For this reason, we ask the question whether *a Fay factor of 10% or 30%* would improve the results. Notably, while the *Fay factor* is commonly applied to *BRR* to address *the missing replicate challenge* (as described earlier), this issue is not typically associated with the *JK2* method. Nonetheless, this study examines the effect of introducing a *Fay factor* (*10%*, *30%*, and *50%*, respectively) to *JK2 Half* on the estimation of sampling variance, particularly for non-smooth statistics.[12] Our comparisons across different variance estimators are also useful to evaluate whether the increased computational effort required by *JK2 Full* is justified by the precision gained in the sampling error of the variance estimator.

Our additional motivation in this research is to examine the effects of varying conditions of sample sizes and non-response. Valliant et al. (2018) suggested that

[12] We investigated the effect of introducing *Fay* factors (*10%*, *30%*, and *50%*) to the *JK2 Full* method. However, as the results were very similar to those of *JK2 Half*, we opted not to report them here for the sake of simplicity and clarity in presenting the findings of this report.

collapsing variance strata or adopting a *one-PSU-per-variance-stratum* approach often leads to overestimation of sampling variance in replication methods. This can occur, for example, in scenarios where there is *an odd number of PSUs* in a given explicit stratum. As described in the "*Odd Number of PSUs in the explicit stratum*" section, *BRR* addresses this issue by collapsing two variance strata into one, forming triplets of PSUs within a single variance stratum. We hypothesize that this approach may lead *BRR* to overestimate the sampling variance based on Valliant et al.'s (2018) findings. This overestimation effect may be more pronounced in smaller sample sizes, as larger sample sizes are likely to mitigate this issue to some extent. Similarly, Valliant et al. (2018) indicated that if sub-PSU units (e.g., classes or students in ILSAs) are paired instead of PSUs in a given variance stratum, replication methods are likely to underestimate sampling variance. Based on this, we hypothesize that in such scenarios, *JK2* may underestimate the sampling variance.

As explained in the *"**Non-Response**"* section, one common approach is to leave *PSUs* in their original variance strata, referred to in this paper as the *OrgStr* approach, or to pair only participating PSUs. Under the *OrgStr* approach, *BRR* treats each participating PSU as a single PSU within its original variance stratum. For example, if eight schools are sampled and four do not participate, the result could be four variance strata, each containing a single PSU. In this example, using the *BRR OrgStr* approach, replicate weights are generated based on a single PSU's replicate weight across *R* iterations from each variance stratum. This PSU's weight is alternately inflated or deflated according to *BRREP* procedures, the *Hadamard* matrix combinations, and the chosen *Fay* factor. As a result, within each variance stratum, only one PSU contributes to the variance estimation. In contrast, under the standard *BRR* approach, only participating PSUs are re-paired into new variance strata, allowing two PSUs in each stratum to contribute to the generation of replicate weights—and thus to the variance estimation. Based on the same example, this would result in two (instead of four) variance strata for *R* replicates, with the contribution of each PSU modified by the *BRREP* and Hadamard matrix combinations, as well as the chosen *Fay* factor. Unlike the *OrgStr* approach, standard *BRR* ensures that at least two PSUs contribute to the variance estimation within each stratum.

In *JK2*, traditionally in the ILSAs, there has been no *OrgStr* approach, as *JK2* may not handle this structure effectively. However, for the purpose of a complete comparison between *BRR* and *JK2*, we implement an *OrgStr* approach in both *JK2 Half* and *Full*. In *JK2*, the *OrgStr* approach involves pairing sub-PSU units—such as students or classes—

within a given variance stratum. As a result, some variance strata may consist entirely of sub-PSU units rather than full PSUs. For instance, in the case where eight schools are sampled and four do not participate, each affected variance stratum would include only one PSU, which is then subdivided into quasi sub-PSU units (e.g., classes or students) to derive *R* replicate weights. This approach is applied in *JK2* within ILSAs under two specific conditions: (i) when a school is sampled with certainty, classes or students are treated as paired sub-PSU units; and (ii) when there is an odd number of schools in a given explicit stratum, the remaining unpaired ("orphan") PSU is divided into sub-PSU units and paired accordingly. We hypothesize that this adjustment may lead *JK2* to underestimate sampling variance, as students within the same school (serving as the PSU in ILSA data) are more homogeneous than students across different schools. While this issue may have a minor effect when only one variance stratum requires such adjustments (i.e., *Odd Number of Schools* in a given explicit stratum), the risk of underestimation increases when multiple unpaired school variance strata emerge due to non-responding PSUs. Therefore, readers should interpret the results from the *JK2 OrgStr* approaches with caution. We acknowledge that *JK2 OrgStr* and *BRR OrgStr* handle variance strata differently, making direct comparisons challenging; however, we included both approaches in this study to reflect the current treatment of PSUs in ILSAs. Specifically, *JK2 OrgStr* extends the handling of single-PSU variance strata (as applied in TIMSS [Fishbein et al., 2024]) to account for non-response cases, while *BRR OrgStr* manages non-response scenarios in line with the treatment of PSUs in PISA (OECD, 2024). By systematically investigating these configurations, this study aims to deepen our understanding of how varying variance strata approaches impact variance estimation in ILSAs.

Specifically, we aim to address the following research questions to compare the replication methods:

1. Among *JK2 (Half and Full)*, *BRR* (with and without a *Fay* factor), and *Bootstrapping*, which method performs best for estimating sampling variance in smooth and non-smooth statistics?
2. Does introducing a *Fay* factor (i.e., *10%*, *30%*, and *50%*) in *JK2* and *BRR* improve the precision of sampling variance estimation?
3. How do the examined replication methods respond to varying score distributions (two score variables with ICCs of 25% and 50%), different sample sizes of PSUs, and the introduction of non-response?

4. How does the handling of non-responding PSU's (i.e., *OrgStr* or *adjacent PSU pairing*) influence the performance of the replication methods in estimating sampling variance?
5. Can *Bootstrapping* be effectively applied to ILSA data for variance estimation?

With these research questions, our study aims to contribute to literature in four significant ways. First, while replication methods have been analyzed in the survey methodology literature, their performance has not been compared in the particular context of ILSAs. Second, by evaluating these methods, the study provides practical insights for principal investigators and researchers relying on these methods. Third, the study systematically examines the impact of different conditions, including score distributions, sampling methods, and non-response, to determine their influence on variance estimation. Fourth, this study introduces and evaluates new approaches that are not commonly used in ILSAs, such as applying *Fay's Method* to *JK2* (with *10%*, *30%*, and *50%* factors), testing alternative *Fay* factors for *BRR* (*10%* and *30%*), and assessing *Bootstrapping* as a potential alternative. By doing so, this research should provide valuable guidance for the effective application of replication methods within the context of ILSA data.

5 Methods

Our study examined the performance of variance estimators under different conditions for various types of statistics using a Monte Carlo simulation in R (R Core Team, 2022). Rather than generating and analyzing simulated data in each replication, this study created a large population-level dataset. From this population, 1,000 samples were drawn for each condition using sampling methods commonly employed in ILSAs (see footnote 5). The simulation approach mirrors standard practices in ILSAs, where participating countries provide a comprehensive list of schools and their sampling attributes for the target population (LaRoche et al., 2020). The sampling team selects schools from this defined population, followed by class- and student-level sampling. After data collection, weights and replicate weights are applied to correct for uneven selection probabilities and non-response, enabling the estimation of both the statistics of interest and the sampling variance of the estimates.

5.1 Simulated Population

We simulated a finite population of 5,000 schools and 500,857 students to analyze the performance of sampling variance estimators. This population is designed to reflect the average population size observed in ILSAs, such as the TIMSS Grade 4 population (Siegel & Foy, 2024). Also, the population is designed to minimize the likelihood of selecting certainty schools in our simulation, where schools have a probability of selection equal to one. Otherwise, we would need to sample the same certainty schools in every simulated sample. This approach also allows us to conduct the study without engaging in discussions about the finite population correction factor, given that the sample size is relatively smaller than the population size. The number of students in each school is randomly assigned from a uniform distribution in the interval [1, 200], creating variation in the measure of size (MOS) used to sample schools. The distribution results in a symmetrical distribution of student enrollment with an average school having around 100 students. We also summarized population characteristics of simulation into **Table 2**.

Table 2: Population Characteristics in Simulation

Variable	Group	Value
Score 1 Mean (SD)	Overall	500 (80)
Score 1 Mean (SD)	Girls	504 (80)
Score 1 Mean (SD)	Boys	496 (80)
Score 1 25th, 50th, 75th Percentiles	Overall	446, 499, 554
Score 1	ICC	0.25
Score 2 Mean (SD)	Overall	500 (80)
Score 2 Mean (SD)	Girls	513 (79)
Score 2 Mean (SD)	Boys	487 (79)
Score 1 25th, 50th, 75th Percentiles	Overall	446, 499, 554
Score 2	ICC	0.50
Correlation Statistic	Score 1 and Score 2	0.35
Number of Schools	Overall	5,000
Number of Students	Overall	500, 857
Number of Students	Girls	250, 428
Number of Students	Boys	250, 429

Student achievement scores are modeled using the following linear equations[13]:

For **Score 1**, the equation is:

$$\text{Score 1} = 500 + (\beta_1 \text{W}_1) + (\beta_2 \text{Private}) + (\beta_3 W2) + (\beta_4 X1) + E_1 \quad (13)$$

where $E_1 \sim N(0,6.8)$, corresponding to an ICC of 25% on school-effects.

For **Score 2**, the equation is:

[13] The simulation procedure closely follows the approach outlined by Atasever et al. (2025), who examined the use of weights in multilevel models within the context of ILSAs.

$$\text{Score 2} = 500 + (\beta_1 \text{W}_1) + (\beta_2 \text{Private}) + (\beta_3 W2) + (\beta_4 X1) + E_2 \quad (14)$$

where $E_2 \sim N(0,4)$, corresponding to an ICC of 50% on school-effects.

All β coefficients were set to 3, except for β_4 which is set to 1, and the intercept was set to 500. Three school level variables were generated. W_1 follows a normal distribution, $N(0,1)$; $W2$ is a binomial distribution with probabilities of 1, indicating whether a school has 100 or more students. A binary stratification variable, Private (i.e., $private$ [1] & $public$ [0]) was generated using a binomial distribution where the probability of being private was determined by the logistic transformation of $W1 - 1.45$. As a result, the simulated population consists of 76% public schools and 24% private schools. By incorporating the Private variable into the equations, this approach models private schools as more likely to achieve higher average outcomes for both Score 1 and Score 2 compared to public schools. This allows us to see how using school type as an explicit stratification affects the performance of variance estimators. The PPS sampling also plays a role because the score variables depend on factors like school size ($W2$) and whether the school is private or public. At the same time, not all variation is explained by the sampling attributes. A school-level error term was added to the linear equation, as $E_1 \sim N(0,6.8)$ for an ICC of 25% and $E_2 \sim N(0,4)$ for an ICC of 50%. The setup of the simulated population and variables is designed to reflect realistic variation in achievement with random effects ($W1, X1, and\ E_2$) while allowing us to assess how explicit stratification ($Private$) and PPS sampling ($W2$) explain this variation. The Pearson correlation coefficient between Score 1 and Score 2 is 0.35 (see **Table 2**).

Within each school, students were randomly distributed into classrooms with an average size of 20 students. If a class contained fewer than 10 students (half the average size), those students were reassigned to the first class in the school. As a result, classroom sizes ranged from 10 to 29 students, except in very small schools where all students were placed in a single classroom. Schools had between 1 and 10 classrooms, depending on their total enrollment. To distribute scores within schools at the student level, a single variable $X1$ was created which followed a standard normal distribution, $N(0,1)$.

Achievement scores generated based on equations (13) and (14) above were designed to reflect the typical TIMSS country metric scale. These scores were re-scaled to follow a multivariate normal distribution with a mean of 500 and a standard deviation (SD) of 80, while preserving intraclass correlations (ICCs) of 25% and 50%. These scores

align with the average ICC and the maximum ICC observed in TIMSS (Foy & Joncas, 2004) and account for clustering effects at the school level. Additionally, the average mathematics mean score for TIMSS 2023 was calculated as 504, with a standard deviation of 84 and an average school-level ICC of 26%, while the maximum ICC reached 50% (see **Table Appendix A 1** for details).

For this study, an SD of 80 was chosen to reflect the average SD observed in TIMSS 2023 (e.g., see TIMSS 2023 average in the **Table Appendix A 1**). However, ICC plays a more critical role in influencing sampling variance. An ICC of 25% indicates that a quarter of the variance in student achievement in a population is attributable to school-level effects, representing the average clustering effect in TIMSS data. In contrast, an ICC of 50% suggests that half of the variance in student achievement is explained by differences between schools, highlighting a stronger school-level effect on achievement outcomes.

To create a binary gender variable ($girls$ [1] & $boys$[0]), students were first sorted in descending order based on their Score 2 results. To ensure a structured gender distribution aligned with typical achievement patterns, boys and girls were assigned to the top and bottom halves based on their achievement levels. Specifically, 60% of the girls were placed in the top half, while the remaining 40% were in the bottom half. The remaining portion of the top half was filled with boys. Similarly, the lower half contained the remaining boys and girls. Following this stratification, students were randomly assigned a binary variable ($girls$ [1] & $boys$[0]) from the sorted list, ensuring that, on average, girls outperformed boys in achievement. The simulated population maintained equal gender distribution, with 50% boys and 50% girls. The population descriptive statistics indicate that the mean Score 1 for boys is 496 ($SD = 80$) while for girls, it is 504 ($SD = 80$), and the mean Score 2 for boys is 487 ($SD = 79$) while for girls, it is 513 ($SD = 79$). The gender variable allows for structured comparisons in mean analyses and significance testing as a grouping variable (see **Table 2**).

5.2 Sampling Design and Weights

We implemented a two-stage sampling design that mirrors the standard sampling process in TIMSS (Siegel & Foy, 2024). We summarize the characteristics of samples in our simulation into **Table 3.**

Schools are sampled using the systematic PPS method, with implicit stratification achieved by sorting schools by size (large to small) within each explicit stratum. Explicit stratification was based on school type, with a binary classification of $private\ (1)\ and\ public\ (0)$. Schools were sampled in proportion to their representation in the population—24% private and 76% public. For instance, in a sample of 100 schools, 24 private schools and 76 public schools were selected.

The probability of selecting a school (π_i) in a given explicit stratum is given by:

$$\pi_i = \frac{nS_i}{\sum S_i} \tag{15}$$

where n is the number of schools to be sampled, S_i is the number of students in school i, $\sum S_i$ represents the total student population (N) within a given explicit stratum (see Siegel & Foy, 2024, Appendix 3A).

A systematic PPS approach was applied by first sorting schools by size and calculating the sampling interval $\sum S_i/n$. (Siegel & Foy, 2024). A random number from a uniform distribution [0,1] was drawn and multiplied by the sampling interval to determine the first selected school. Subsequent schools were selected sequentially based on cumulative MOS, each with the same sampling interval (see Siegel & Foy, 2024 Appendix 3A). The school weight (*schwgt*) in a given explicit stratum is the inverse of the selection probability:

$$schwgt_i = \frac{1}{\pi_i} \tag{16}$$

Table 3: Characteristics of Samples in Simulation

Category	Details
Simulation Iterations	1,000 for each scenario
Explicit Stratification	Schools explicitly stratified and proportionally distributed as Private (24%) and Public (76%)
Implicit Stratification	Schools sorted by size (large to small) within each explicit stratum. No other implicit stratification is used.
School Sampling Method	Systematic with Probability Proportional to Size (PPS)
Very Small School Samples	29 (22 Public and 7 Private), 30 (22 Public and 8 Private)
Small School Samples	49 (38 Public and 11 Private), 50 (38 Public and 12 Private)
Medium School Samples	99 (76 Public and 23 Private), 100 (76 Public and 24 Private)
Large School Samples	149 (114 Public and 35 Private), 150 (114 Public and 36 Private)
School Samples with Non-Response	Initially Sampled: 30, 50, 100, 150 Participating: 26 (4 Private), 44 (6 Private), 88 (12 Private), 132 (18 Private)
Non-participating Schools	~50% of Private schools sampled were randomly selected to be refusals while all public schools were set to be participating.
Non-response Rates	Non-response rates across simulations: ~87% to 88%
Class Sampling Scenario	2 classes in larger schools ($n > 140$), otherwise 1 class
Student Sample Sizes	~900 students for 30 schools ~1,500 students for 50 schools ~3,000 students for 100 schools ~4,500 students for 150 schools

To account for different sampling scenarios, school sample sizes were categorized into very small, small, medium, and large groups. Small samples (private schools) included 29 (24 public and 5 private) and 30 (24 public and 6 private), and very small samples are 49 (38 public and 11 private), and 50 (38 public and 12 private) schools, while medium samples consisted of 99 (76 public and 23 private) and 100 (76 public and 24 private) schools. Large samples included 149 (114 public and 35 private) and 150 (114 public and 36 private) schools (see **Table 3**). In ILSAs, the standard minimum school sample size is 150 (Siegel & Foy, 2024). However, in some countries, the achieved sample may be closer to 100 schools due to a smaller population size or school non-participation (see Siegel et al., 2024). Conversely, small school samples (e.g., 30 or 50 schools) may be used for subgroup comparisons, where a specific portion of the population is analyzed based on PSUs. For sample sizes with an *odd number of schools* in a given explicit stratum, one fewer private school ($n - 1$) was selected, slightly reducing the proportional representation of private schools. While smaller samples (30–50 schools) are commonly used for subgroup analyses in ILSAs, larger samples (150 schools) represent full-population sampling strategies.

Non-response was applied to the simulations to reflect real-world conditions in ILSAs, where non-response patterns often differ among strata, e.g., private schools are typically more likely to refuse participation compared to public schools. To simulate a significant non-response scenario, 50% of private schools were randomly designated as non-participating by selecting every second school from a list of sampled schools, while all public schools were set to participate. For initial school samples of 30, 50, 100, and 150 schools, the number of participating schools was reduced to 26 (22 public and 4 private), 44 (38 public and 6 private), 88 (76 public and 12 private), and 132 (114 public and 18 private), respectively, resulting in non-response rates of approximately 12% to 13% (see **Table 3**). To adjust for non-participation, a participation adjustment factor was applied to the school weights. For instance, if an explicit stratum originally included 24 sampled schools, but 12 of them did not respond, the non-response adjustment factor was calculated as:

$$schad_i = \frac{24}{24 - 12} = 2 \tag{17}$$

The final school weight is calculated as, *schwgt* × *schadj*. $Schadj$ equals to 1 in sampling scenarios, where no non-response is present.

At the second stage of our sampling procedure, within the selected schools, two classes were sampled in larger schools ($S_j > 140$), and one class is selected in smaller schools. If the school sample size was 150, student sample size approximately consisted of 4,500 students (i.e., a typical student sample size for TIMSS).

The conditional class weight ($clswgt$) was calculated based on the number of classrooms within a school. If only one class was selected, $clswgt$ equaled the total number of classrooms. If two classes were selected, $clswgt$ was computed as the total number of classrooms divided by two, i.e., the number of selected classes.

Scenario 1: in a school with 100 students and 5 classrooms, one classroom is randomly sampled, and all students in that class have a weight of:

$$clswgt = 5, \quad \text{since } 5 \times 20 = 100 \tag{18}$$

Scenario 2: in a school with 180 students and 5 classrooms, one classroom is randomly sampled, and all students in that class have a weight of:

$$clswgt = \frac{9}{2} = 4.5, \quad \text{since } 4.5 \times 2 \times 20 = 180 \tag{19}$$

The total weight for each student is computed as:

$$totwgt = schwgt \times schadj \times clswgt \times stdwgt \tag{20}$$

where $stdwgt = 1$, since student non-participation was not simulated, and no additional direct sampling occurred at the student level. $totwgt$ refers to the final student weight, which is used to estimate the statistics of interest. This same weight variable was also applied in generating replicate weights to estimate the sampling variance.

5.3 Variance Estimators and Replicate Weights

We evaluated variance estimators categorized into three main replication methods: *JK2*, *BRR*, and *Bootstrapping*.

5.3.1 Jackknife Repeated Replication (JK2)

For the *Jackknife* method, we analyzed *JK2* in both *Full* and *Half* variants, as previously described. Additionally, we explored the potential application of *Fay* variants for the *JK2* method, which has not been commonly implemented in ILSAs. A *Fay* factor

(f), where $0 \leq f < 1$, can be applied to adjust *JK2*'s replicate weight factors in both *Full* and *Half* methods.

For example, if $f = 0.5$, the Jackknife replicate adjustment factors are modified as follows:

$$rw_r = \begin{cases} 1.5 \cdot w_1, if\ JKREP = 1 \\ 0.5 \cdot w_2, if\ JKREP = 0 \end{cases} \quad (21)$$

We also analyzed the *Fay* factors $f = 0.1$ and $f = 0.3$ for *JK2* estimators. If a *Fay* factor is applied to *JK2 Full*, the MSE is multiplied by a scaling factor (k):

$$k = \frac{1}{2 * (1 - f)^2} \quad (22)$$

For example, if f = 0.5, then: $k = \frac{1}{2*(1-0.5)^2} = 2$. In the *JK2 Half* with *Fay* of 50%, $k = \frac{1}{(1-0.5)^2} = 4$. This scaling factor ($k = \frac{1}{(1-f)^2}$) was originally introduced by *Fay* (1984) for *BRR* to adjust for the effect of the *Fay* adjustment parameter f on variance estimation. Given the similar mathematical structure and replication logic underlying *JK2*, we propose that the same scaling factor is equally applicable to *JK2*. While *BRR* requires an additional division by $\frac{1}{R}$, this adjustment is unnecessary for the *JK2*.

In addition to the traditional *JK2 Full* and *JK2 Half* estimators, we examined the impact of preserving the original variance strata structure on the precision of sampling variance estimation. Rather than redefining variance strata to account for PSU non-response, the *JK2 Full Original Strata* (*OrgStr*) and *JK2 Half Original Strata* (*OrgStr*) methods retain the initial variance strata structure as defined at the sampling stage. This allows for a direct comparison with traditional *JK2 Full* and *JK2 Half* estimators to evaluate whether variance strata adjustments impact variance estimation precision.

5.3.2 Balanced Repeated Replication (BRR)

We examined *BRR* without a *Fay* factor alongside *Fay*-adjusted variants with *Fay* factors of 10% ($f = 0.1$), 30% $f = 0.3$, and 50% $f = 0.5$. While traditionally in ILSAs, the *BRR* and *BRR* with $f = 0.5$ are widely used, we aimed to assess whether the precision of the estimators improves when applying lower *Fay* factors $f = 0.3$ and $f = 0.1$. For all methods employing *Fay*'s adjustment, the scaling factor k is defined as follows, where R represents the number of replicate weights:

$$\frac{1}{R * (1-f)^2} \tag{23}$$

In standard *BRR* without a *Fay* factor, the scaling factor simplifies to: $\frac{1}{R}$.

For non-response scenarios, we introduced *BRR* with 50% Fay under the non-response scenario with *Original Strata* (*OrgStr*) approach. Similar to *JK2 OrgStr*, this method retains the original variance strata, ensuring that only responding PSUs remain within their respective variance strata, without re-pairing them with other participating PSUs. This allows for evaluating the impact of maintaining the original sampling structure on the precision of variance estimation under conditions of PSU non-response.

For each *BRR* method, an orthogonal Hadamard matrix is generated, where H is the smallest multiple of 4 that is greater than or equal to the number of sampled schools S. If S is already a multiple of 4 (e.g., $S = 100$), then $H = S$. If S is not a multiple of 4 (e.g., $S = 30$), then H is rounded up to the nearest multiple of 4 (e.g., $H = 32$).

For both *JK2* and *BRR* methods, we ran scripts in R (R Core Team, 2022) that work with the IEA's IDB Analyzer (IEA, 2025).

5.3.3 Bootstrapping

In addition to the *JK2* and *BRR* methods, we implemented *Bootstrapping* specifically designed for cluster sampling in ILSA. This method was executed using R (R Core Team, 2022) with *svrep* (Schneider, 2023) and *survey* packages (Lumley, 2004) to align with the ILSA framework. In each variance stratum with a fixed sample size of n sampling units, each replicate resamples $(n-1)$ units with replacement. For *Bootstrapping*, we consistently used the variance strata defined by the *JK2* design, treating paired schools as PSUs. For each replicate, two unique replicate weights were derived at random using the Bootstrapping algorithm. Unlike *JK2*, where weights are deterministically adjusted by removing one PSU at a time, the *Bootstrapping* method randomly assigns one PSU to have its weight inflated and the other deflated—in a manner conceptually similar to *BRR*. However, instead of using a *Hadamard* matrix as in *BRR*, the inflation/deflation in *Bootstrapping* is determined randomly for each replicate.

In this sense, *Bootstrapping* can be viewed as conceptually falling somewhere between *BRR with Fay* adjustment and *JK2*: it allows flexible replication, where each replicate involves randomly adjusting the weights of PSU pairs. The number of replicates

can be defined by the user, and in each replicate, one PSU's weight is increased while the other's is decreased, preserving the overall structure of the design. The methodology of inflating or deflating the weights closely follows *Fay's Method* (i.e., keeping all PSUs) but with one key distinction: replicate weight adjustment factors were computed based on the total final weight, allowing each PSU's adjustment factor to vary across replicates. However, these factors typically fall into two categories: one PSU's weight is inflated, while the other is deflated based on a randomized assignment. For further details on this *Bootstrapping* approach, refer to Beaumont and Émond (2022).

An example of replicate weight adjustment factors is computed for a single variance stratum, with this process systematically applied across all variance strata. For example, let variance stratum h have two sampled PSUs, so that $n_h = 2$. The number of PSUs contributing to the variance estimate within the stratum is as follows, $m_h = n_h - 1 = 1$. The number of replicate weights generated per variance stratum is given by $\mathrm{r} = 2$.

The multiplicities (i.e., the number of times each PSU is resampled) follow a multinomial random distribution, which determines whether a PSU's weight is inflated or deflated in a given replicate.:

$$\text{multiplicities} \sim \text{Multinomial}(n = \mathrm{r}, \text{size} = m_h, \text{prob} = (n_h)^{-1}) \tag{24}$$

For a given variance stratum, this results in the following assignment:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{25}$$

Each PSU is assigned a replication factor based on:

$$a_k = 1 - \sqrt{\frac{m_h(1-\pi_k)}{n_h - 1}} + \sqrt{\frac{m_h(1-\pi_k)}{n_h - 1}} \times \left(\frac{n_h}{m_h}\right) \times m_k^* \tag{26}$$

Where π_k is the probability weight of a student, defined as $\pi_k = \frac{1}{\text{totwgt}}$. And m_k^* represents the resampling indicator from the multinomial process. Assuming the total final weight is $\text{totwgt} = 200$, then the probability is $\pi_k = \frac{1}{200} = 0.005$. Then, this results in the following replication factors:

$$a_k = \begin{bmatrix} 0.002503133 & 1.997496867 \\ 1.997496867 & 0.002503133 \end{bmatrix} \tag{27}$$

The replicate weights for the PSUs in the first and second replicate, for a single variance stratum, are given by:

$$rw_r = \begin{cases} 0.002503133 \cdot w_1, 1.997496867 \cdot w_1 \\ 1.997496867 \cdot w_2, 0.002503133 \cdot w_2 \end{cases} \quad (28)$$

This process is repeated across all variance strata. Finally, to estimate the sampling error, *Equation 1* is scaled by an adjustment factor $\mathrm{k} = \frac{1}{\mathrm{R}}$ where R represents the total number of replicate weights used in the variance estimation.[14] This makes the formula of *Bootstrapping* equivalent to the standard *BRR*.

5.4 Number of Replicate Weights

For *JK2 Half*, *BRR*, and their *Fay* variants, the number of replicate weights equals the number of variance strata, calculated as:

$$\text{Variance Strata} = \frac{S}{2}, \quad \text{Replicate Weights} = \frac{S}{2} \quad (29)$$

For example, if 100 schools are sampled, Variance Strata $= 50$, Replicate Weights $= 50$.

For *JK2 Full*, the number of variance strata remains the same, but the replicate weights equal the number of sampled schools:

$$\text{Variance Strata} = \frac{S}{2}, \quad \text{Replicate Weights} = S \quad (30)$$

For example, if 100 schools are sampled, Variance Strata $= 50$, Replicate Weights $= 100$. For *Bootstrapping*, the number of replicate weights follows the *JK2 Full* approach, Replicate Weights $= S$. Thus, if 100 schools are sampled, 100 replicate weights are created.

If non-response is simulated, the number of variance strata is based on the number of participating schools (P):

[14] Variance estimation for TIMSS data using *Bootstrapping* can be computed with the survey and svrep packages in R. The survey design is specified with stratification and PPS sampling using svydesign(), where *IDSCHOOL* are defined as Primary Sampling Units (PSUs) and the *JKZ* variable as the stratum. The design is then converted into a Rao-Wu-Yue-Beaumont *Bootstrap* design via as_*bootstrap*_design() from svrep. Weighted means and standard errors are estimated using svymean(), incorporating replicate weights to estimate sampling variance.

$$\text{Variance Strata} = \frac{P}{2}, \quad \text{Replicate Weights} = \frac{P}{2} \tag{31}$$

For instance, if 88 schools participate, Variance Strata $= 44$, Replicate Weights $= 44$. For *BRR OrgStr* and *JK2 Half OrgStr*, the variance strata are kept at their original number, but replicate weights are based on the number of participating schools:

$$\text{Variance Strata} = \frac{S}{2}, \quad \text{Replicate Weights} = \frac{P}{2} \tag{32}$$

For *JK2 Full OrgStr*, the variance strata remain unchanged, but the replicate weights equal the number of participating schools:

$$\text{Variance Strata} = \frac{S}{2}, \quad \text{Replicate Weights} = P \tag{33}$$

For instance, if 100 schools were originally sampled but only 88 participated. *BRR OrgStr* and *JK2 Half OrgStr*: 50 variance strata, 44 replicate weights. - *JK2 Full OrgStr*: 50 variance strata, 88 replicate weights. The example referring to a sample of 24 private schools where only 12 of them participated would result in a scenario as follows: - In the *BRR OrgStr* approach, for private schools, 12 variance strata are retained, each contributing a single PSU to the variance estimation. This contrasts with the standard *BRR* approach, where at least two PSUs contribute to the variance estimation within each variance stratum. In the *JK2 OrgStr* approach, the same 12 variance strata are maintained for private schools. However, instead of leaving one PSU in a given variance stratum, each PSU is internally split into sub-PSU units—such as classes or students—which are then randomly assigned to a *JKREP* number to form 24 sub-PSU pairings.

The *BRR OrgStr* approach does not exactly replicate the methodology used in PISA. In ILSAs that use *BRR* such as TALIS and PISA, 100 replicate weights are computed regardless of the number of variance strata. For instance, a country may have 80 or 100 variance strata, but in either case, 100 replicate weights are generated. In our approach, we chose to base the number of replicate weights on the number of participating schools (P). Our rationale for this decision is rooted in the need to compare different variance estimation methods—namely, *JK2* (and its variants), *BRR* (and its variants), and *Bootstrapping*—on an equal footing. To ensure comparability, we standardized the number of replicate weights across methods.

Only *JK2 Full* and *Bootstrapping* involve twice the number of replicates, as they derive two replicate weights per variance stratum by design. However, the treatment of single-PSU strata follows the PISA approach in the *BRR OrgStr* method, where replicate weights can still be derived from a single PSU if the other PSU is missing due to non-response. In contrast, standard BRR and its variants derive replicate weights from two PSUs within each variance stratum, as the participating PSUs are re-paired into variance strata, as in TALIS. The **Table 4** below summarizes the sampling methods, the corresponding number of originally sampled schools (S), variance strata (VS), the number of replicate weights (R), and the number of participating schools (P).

Table 4: Variance Estimators Examined in the Study

Variance Estimator	Number of PSUs S	Variance Strata VS	Number of Replicates R
Full Participation Scenario			
JK2 Half	S	$S/2$	$S/2$
Fay's Method for JK2 Half (10%, 30%, 50%)	S	$S/2$	$S/2$
BRR	S	$S/2$	$S/2$
Fay's Method for BRR (10%, 30%, 50%)	S	$S/2$	$S/2$
JK2 Full	S	$S/2$	S
Bootstrapping	S	$S/2$	S
Non-response Scenario			
JK2 Half	P	$P/2$	$P/2$
Fay's Method for JK2 Half (10%, 30%, 50%)	P	$P/2$	$P/2$
BRR	P	$P/2$	$P/2$
Fay's Method for BRR (10%, 30%, 50%)	P	$P/2$	$P/2$
JK2 Full	P	$P/2$	P
Bootstrapping	P	$P/2$	$P/2$
BRR OrgStr	P	$S/2$	$P/2$
JK2 Half OrgStr	P	$S/2$	$P/2$
JK2 Full OrgStr	P	$S/2$	P

Note: when there is 100% response rate simulated, $S = P$.

5.5 Estimated Parameters of Interest

The Horvitz–Thompson (HT) estimator is widely used in survey sampling and provides unbiased estimates for both linear and non-linear statistics, assuming that no element

(e.g., students) is sampled with replacement (Mang et al., 2021). The general form of the HT estimator is:

$$\hat{y} = \frac{1}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathrm{w_{ij}}} \sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathrm{w_{ij}} y_{ij} \tag{34}$$

where: $\hat{y}$ is the estimated statistic, y_{ij} represents the observed value (or score) from the j-th student in the i-th school, w_{ij} is the inverse of the selection probability of student j in school i, m is the number of sampled schools, and n_i is the number of students sampled within school i (Horvitz & Thompson, 1952; Mang et al., 2021). The HT estimator is applied to compute key population parameters of interest. For each selected sample, we estimate a set of population parameters that are widely used in educational policy and commonly reported in ILSAs. This study examines both smooth and non-smooth statistics for Score 1 and Score 2. Specifically, the parameters of interest include means, differences in means, percentiles, benchmark percentages, and correlations.:

1. For means, we conducted analyses at both the overall level and sub-group level (e.g., by gender), with sub-group analysis facilitating the calculation of differences in means.
2. Percentiles, such as the median (50th percentile), 25th percentile, and 75th percentile, were estimated using cumulative weighted distributions at both the overall level and the sub-group level (e.g., by gender).
3. Benchmark percentages, which represent the proportion of students meeting or exceeding specific achievement thresholds (400 for low,[15] 550 for high, and 625 for advanced performance), were also estimated at the overall level.
4. Additionally, Pearson correlations between Score 1 and Score 2 were estimated at the overall level.

The parameter estimation algorithms used in this study are identical to those applied in standard statistical software designed for complex survey designs, such as the IEA IDB Analyzer (IEA, 2025). For further information how these statistics are computed, please see *IEA's IDB Analyzer* (IEA, 2025) and *Survey* package help manuals (Lumley, 2004).

[15] We also analyzed the results for the intermediate benchmark (475); however, we chose not to report them, as they closely mirrored the findings for the low benchmark.

5.6 *Relative Bias* and *Stability* of Variance Estimators

We summarized the results of the study simulation as follows. For each sample and a given population parameter of interest, we estimated the sampling variance using different variance estimation methods. The sampling error is then computed as the squared root of the sampling variance.

$$se(\hat{\epsilon})^t = \sqrt{var(\hat{\epsilon})^t} \tag{35}$$

The population parameter of interest is $\hat{\epsilon}$ and the sampling variance is $var(\hat{\epsilon})^t$. t represents the respective variance estimation method listed in **Table 4.**

The final average estimate of the sampling error is computed as the average of sampling error estimates across 1,000 iterations.[16] The following equation indicates the average sampling error:

$$\overline{se}(\hat{\epsilon})^t = \frac{\sum_{i=1}^{1000} se(\hat{\epsilon})^t}{1000} \tag{36}$$

The "*true*" sampling error is computed by measuring the standard deviation of the parameter estimates across all 1,000 samples in the simulation as an approximation of the standard deviation of the *Full* sampling distribution (Meinck & Vandenplas, 2021):

$$se(\hat{\epsilon})^* = \sqrt{\frac{\sum_{i=1}^{1000}(\hat{\epsilon}_i - \hat{\epsilon}^*)^2}{1000 - 1}} \tag{37}$$

The *Relative Bias* is calculated as the ratio of the square root of the average sampling variance estimated by the methods under investigation to the standard deviation of the sampling distribution of the parameters of interest, i.e., the "*true*" sampling error:

$$Relative\ Bias_{\overline{se}(\hat{\epsilon})^t} = \frac{\overline{se}(\hat{\epsilon})^t}{se(\hat{\epsilon})^*} \tag{38}$$

$\hat{\epsilon}^*$ is the mean over $\hat{\epsilon}$ across all 1,000 samples in the simulation study.

[16] Given our goal of comparing the performance of variance estimators in capturing sampling error based on a sampling distribution of 1,000 samples, we compare the average of the square roots of the sampling variances, rather than the square root of the average sampling variance.

The *Stability* of the sampling error is calculated as the ratio of the standard deviation of the sampling errors estimated by the methods under investigation to the standard deviation of the sampling distribution of the parameters of interest, i.e., the "*true*" sampling error:

$$Stability_{\overline{se}(\hat{\epsilon})^t} = \frac{sd\,(se(\hat{\epsilon})^t)}{se(\hat{\epsilon})^*} \tag{39}$$

For *Relative Bias* and *Stability*, we follow the simulation methodology outlined by Judkins (1990), although we additionally use the standard deviation around the estimate and the variance of the estimate to capture the distribution around the mean of the estimates. This interpretation of *Relative Bias* is also applied in other ILSAs simulation studies (Mang et al., 2021; Mang et al., 2024). In other simulation studies for replication methods, *Relative Bias* (or simply bias) is commonly expressed in the literature as $\overline{se}(\hat{\epsilon})^t - se(\hat{\epsilon})^*$ or as the absolute value of this difference (See Kovar, 1985; Kovar et al., 1988). However, this absolute measure can fluctuate significantly depending on the type of statistic being analyzed—for example, the SE of a correlation coefficient versus the SE of a median. To enable a meaningful comparison of variance estimators across different statistics, we use *Relative Bias*, which standardizes bias relative to the "*true*" sampling error. For example, if *Relative Bias* = 1, the variance estimator accurately approximates the true sampling error. If *Relative Bias* = 1.10, the estimator overestimates the sampling error by 10 percentage points. If *Relative Bias* = 0.90, the estimator underestimates the sampling error by 10 percentage points. On the other hand, *Stability* reflects the variability of the sampling error relative to the estimate. This measures, in other words, the standard error of the sampling error. Lower *Stability* values indicate a more precise variance estimator. For example, if *Stability* = 0.50, the variance estimator's distribution is half the distribution of the estimate. If *Stability* = 0.10, the variance estimator is more stable and precise. Our approach allows us for a more robust examination of the replication methods across diverse statistical estimators.

6 Results

The results across all tables in this section summarize the performance of variance estimators for the statistical parameters analyzed. These include comparisons of *Relative Bias* and *Stability*, averaged across two score estimates (e.g., Score 1 with 25% ICC and Score 2 with 50% ICC). Each table evaluates results under four sample size scenarios (very

small, small, medium, and large) and three condition scenarios (regular, odd-number PSUs, and non-response).

1. ***Regular Sample Scenario***: Very small (30 schools), small (50 schools), medium (100 schools), and large (150 schools).
2. ***Odd-Number of Schools Scenario***: Very small (29 schools), small (49 schools), medium (99 schools), and large (149 schools).
3. ***Non-Response Scenario***: Very small (26 schools), small (44 schools), medium (88 schools), and large (132 schools).

In each results table per sampling scenario, we highlight the estimators closest to a *Relative Bias* of 1 and those closest to a *Stability* value of 0. We also note that, below each table, the average of the estimated statistics—such as an overall means of 500—is reported. These values are approximately identical to the population values, *Regular Sample*, *Odd-Number of Schools,* and *Non-Response* scenarios, as the estimates based on 1,000 samples are unbiased by definition. However, minor differences between the average of the estimated sample statistics and the population statistics can be observed across sample size scenarios. These differences are attributable to random sampling variation, sample size, and the choice of statistical analysis—particularly when non-smooth statistics are involved. Additionally, the estimated statistics do not differ across the variance estimators, as all methods use the same total student weight for the estimations.

6.1 Means (Overall)

Table 5 compares variance estimators for mean statistics at the overall level, assessing their *Relative Bias* and *Stability* across the three sampling scenarios described earlier. A *Relative Bias* of 1 reflects an unbiased estimator that accurately approximates the true sampling error, while lower *Stability* values indicate higher precision in variance estimation.

[pls insert **Table 5**: Variance Estimator Comparison - Means Statistics (Overall) about here]

In the ***Regular Sample Scenario***, *Relative Bias* is minimal. All estimators, including *BRR* and *JK2* variants, perform similarly for very small samples with only two percentage points of *Relative Bias*. The best performing method across all conditions is *Bootstrapping*, which shows no *Relative Bias* for medium samples. However, in terms of *Stability*, the *JK2* estimators consistently outperform *BRR* and *Bootstrapping*, while *Bootstrapping* has the poorest *Stability*. Other key findings reveal identical performance among *BRR* and its *Fay*

variants, as well as between *JK2 Half* variants. Another notable finding is that variance estimators perform well even with a relatively small number of PSUs, such as 30 or 50 schools. However, when considering the *Stability* indicator, it becomes evident that as the sample size increases, *Stability* decreases. This suggests—as can be expected—reduced variability in the sampling error for larger samples.

Under the ***Odd Number of Schools Scenario***, where one fewer private school is sampled compared to the ***Regular Sample Scenario***, some distinct patterns emerge. In very small samples, *JK2* and standard *BRR* perform best, with only one percentage point *Relative Bias*. For small samples, *Bootstrapping* shows the best performance with three percentage points *Relative Bias*, while *JK2* only exhibits five percentage points *Relative Bias*. In medium samples, *BRR* with *Fay* factors and *JK2* are the best performers, both showing two percentage points *Relative Bias*, with occasional under- and over-estimation. For large samples, *Relative Bias* is similar across estimators, but *Bootstrapping* shows the least *Relative Bias*. Interestingly, *BRR* with a 50% *Fay* factor performs poorly in very small samples, with 13 percentage points *Relative Bias*, and in small samples, with ten percentage points *Relative Bias*. In contrast, the 10% and 30% *Fay* variants perform better, but standard *BRR* outperforms all *Fay* variants. Regarding *Stability*, *JK2* variants consistently show the best indicators across all sample sizes, with no differences between the variants. In small samples, *BRR* with a 50% *Fay* factor performs worse than other *BRR* variants. *Stability* indicator decreases as sample sizes increase, highlighting reduced variability in sampling error for larger samples.

Under the ***Non-Response Scenario***, where school sample sizes range from 26 (very small) to 132 (large), the performance of variance estimators shows clear patterns. The standard *BRR* and its *Fay*-adjusted versions (10%, 30%, and 50%) have a *Relative Bias* of 6 percentage points for very small and small samples, which decreases to 0 percentage points for medium and large samples. *Stability* improves as sample sizes increase, starting at 0.27 for very small samples and dropping to 0.10 for large samples. There are no significant differences between the *Fay*-adjusted and *standard BRR*. *BRR* with *Fay* 50% *OrgStr* shows inflated *Relative Bias*, ranging from 28 to 37 percentage points, although its *Stability* is similar to other *BRR* variants. *Bootstrapping* performs slightly better in terms of *Relative Bias*, showing 3 percentage points for very small samples and 0 percentage points for medium and large samples. However, its *Stability* is weaker, at 0.28 for very small samples and 0.12 for large samples. Both *JK2 Full* and *Half* estimators perform identically, with a *Relative Bias* of 6 percentage points for very small and small samples and 0

percentage points for medium and large samples. *Stability* for these estimators improves from 0.25 for very small samples to 0.10 for large samples. *Fay*-adjusted *JK2* variants show no difference over the standard *JK2*. The *OrgStr* variants of *JK2* differ notably. *JK2 Full OrgStr* underestimates sampling errors considerably, at 42–44 percentage points across all sample sizes, but offers superior *Stability*. *Stability* improves from 0.16 for very small samples to 0.06 for large samples. *JK2 Half OrgStr* shows a *Relative Bias* of 19–22 percentage points, with *Stability* improving from 0.17 for very small samples to 0.07 for large samples.

Across all scenarios, *JK2* estimators consistently demonstrate the best *Stability*, with values decreasing as sample sizes increase. *Relative Bias* is generally minimal for standard *BRR*, *JK2*, and *Bootstrapping*, though *BRR with Fay* 50% *OrgStr* shows inflated sampling errors and *JK2 OrgStr* variants indicate deflated sampling errors. *Stability* is slightly weaker for *Bootstrapping* compared to *JK2*, particularly in smaller sample sizes, but all methods converge in performance as sample sizes grow.

6.2 Sub-Group Means (Gender = Girls)

Table 6 presents the variance estimator comparison for mean estimates of girls comprising a sub-group of interest.[17]

[pls insert **Table 6** about here]

In the ***Regular Sample Scenario***, *Relative Bias* is minimal across all estimators, with values close to 1 for all sample sizes. *BRR* and its *Fay*-adjusted variants, as well as *JK2* estimators, show consistent performance with *Relative Bias* ranging between 1.01 and 1.05. *Bootstrapping* slightly underestimates for very small and small samples (*Relative Bias* of 0.99 and 0.98, respectively) but aligns with the other estimators for larger sample sizes. *Stability* improves with increasing sample sizes across all methods, starting from 0.20–0.23 for very small samples and reducing to 0.08–0.11 for large samples. Notably, *JK2* and *BRR* estimators consistently achieve slightly better *Stability* compared to *Bootstrapping*.

In the ***Odd Number of Schools Scenario***, *BRR* and its *Fay* variants exhibit slightly more variation in *Relative Bias* compared to the Regular Scenario. *BRR* with *Fay* 50%

[17] We also conducted the analysis for boys; however, as the performance of the variance estimators was relatively similar, we report results for girls only.

shows significant overestimation for very small and small samples, with *Relative Bias* values of 12 and 6 percentage points, respectively. Standard *BRR* and *JK2* estimators perform more consistently, with *Relative Bias* values near 1 for most sample sizes. *Bootstrapping* performs similarly to *JK2* for small and large samples but shows slight underestimation for very small samples (*Relative Bias* of 0.97). *Stability* remains comparable across methods, with *JK2* again showing slightly better precision, especially for smaller sample sizes.

Under the ***Non-Response Scenario***, *Relative Bias* is generally higher compared to the previous scenarios. Standard *BRR* and *JK2* estimators show *Relative Bias* values of 6 percentage points for very small and small samples, decreasing to near 0 for medium and large samples. *BRR* with *Fay* 50% *OrgStr* exhibits the most substantial overestimation, with *Relative Bias* values reaching 25–31 percentage points for smaller sample sizes. *Bootstrapping* shows underestimation for medium samples (*Relative Bias* of 0.96) but aligns closer to 1 for other sample sizes. *Stability* indicators reveal that *JK2 OrgStr* variants achieve the highest precision, with *Stability* values as low as 0.06-0.08 for larger sample sizes. However, standard *JK2* and *BRR* estimators show competitive *Stability* across most conditions.

Across all scenarios, *JK2* estimators consistently exhibit superior *Stability* compared to other methods, especially for smaller sample sizes. *Relative Bias* for *BRR* and *JK2* remains close to 1 in most scenarios, indicating high accuracy. *Bootstrapping* shows competitive *Relative Bias* but struggles with *Stability*, particularly for smaller samples. *Fay*-adjusted *BRR* variants generally perform similarly to the standard *BRR*, except for *BRR* with *Fay* 50%, which overestimates sampling error in certain conditions. Notably, *JK2 OrgStr* variants stand out for their precision, particularly under non-response conditions. *Stability* improves across all methods as sample sizes increase, reflecting reduced variability of sampling error estimates in larger samples.

6.3 Differences in Sub-Group Means (Girls-Boys)

Table 7 presents the *Relative Bias* and *Stability* for means differences ($girls - boys$) across different sampling scenarios.

[pls insert **Table 7** about here]

Under the ***Regular Sample Scenario***, all variance estimators show *Relative Bias* values close to 1, indicating high accuracy. *BRR*, *Fay*-adjusted *BRR* variants, and *JK2* estimators

consistently perform well, with *Relative Bias* ranging from 0.97 to 1.00 across all sample sizes. *Bootstrapping* exhibits slightly lower *Relative Bias*, particularly for very small and small samples (0.96 and 0.99, respectively), but remains within an acceptable range. *Stability* improves with increasing sample sizes for all estimators. *JK2* estimators, both *Full* and *Half*, achieve the best *Stability*, decreasing from 0.22–0.23 for very small samples to 0.10–0.12 for large samples. *Bootstrapping* shows the least *Stability*, with values starting at 0.24 for very small samples and reducing to 0.11 for large samples.

In the ***Odd Number of Schools Scenario***, *Relative Bias* values exhibit more variation. Standard *BRR* and *JK2* perform consistently, with *Relative Bias* ranging from 0.95 to 1.00 across all sample sizes. *BRR* with *Fay* 50% shows slightly smaller *Relative Bias* for smaller samples, peaking at 1.00 for very small samples. In this case, *BRR with Fay 50%* shows the best result among the BRR methods. *Bootstrapping* slightly underestimates for smaller samples (0.93 to 0.95) but aligns with other estimators for larger sample sizes. *Stability* follows a similar trend to the Regular Scenario, with *JK2* estimators providing the highest precision. *Stability* values for *JK2* decrease from 0.22 for very small samples to 0.10 for large samples, outperforming *Bootstrapping*, which ranges from 0.24 to 0.12.

The ***Non-Response Scenario*** introduces greater variability in *Relative Bias* and *Stability*. *BRR*, *Fay*-adjusted *BRR* variants, and *JK2* estimators exhibit slight underestimation for small and medium samples, with *Relative Bias* values ranging from 0.96 to 0.99. *Bootstrapping* performs similarly, showing slightly lower *Relative Bias* for medium samples (0.95). *Stability* worsens across all estimators compared to the previous scenarios, with values starting at 0.29-0.31 for very small samples and decreasing to 0.15 for large samples. Notably, *JK2 OrgStr* significantly underestimates *Relative Bias* (0.42–0.45) while achieving the best *Stability*, ranging from 0.14 for very small samples to 0.06 for large samples. *JK2 Half OrgStr* performs slightly better than standard *JK2* in terms of *Stability*, though *Relative Bias* is slightly lower (0.70–0.73).

Across all scenarios, *JK2* estimators consistently perform well in both *Relative Bias* and *Stability*, particularly as sample sizes increase. *BRR* and *Fay*-adjusted *BRR* variants are reliable, though *Fay* 50% shows occasional overestimation for smaller samples. *Bootstrapping* is competitive in *Relative Bias* but struggles with *Stability*, particularly for smaller samples. *JK2 OrgStr* variants stand out for their exceptional *Stability*, despite consistently underestimating the sampling error, as indicated by lower

Relative Bias values. *Stability* improves for all estimators as sample sizes grow, reflecting reduced variability in larger samples.

6.4 Percentiles (25th Percentile, 50th Percentile, and 75th Percentile)

Table 8, **Table 9**, and **Table 10** indicate *Relative Bias* and *Stability* values across percentiles, 25th (the first quarter), 50th (median), and 75th percentiles (the last quarter) at the overall level.[18]

[pls insert **Table 8**, **Table 9** and **Table 10** about here]

In the ***Regular Sample Scenario***, *standard BRR* and *Bootstrapping* generally yield low *Relative Bias* across percentiles, with *BRR* typically exhibiting relative biases around 3 percentage points in smaller samples and approaching 0 to 2 percentage points in larger samples. Specifically, for the 25th percentile, there is a systematic overestimation of approximately 2 percentage points in large samples. In the case of the median, a 4 to 5 percentage point overestimation is observed in medium-sized samples, whereas other estimators show a more modest overestimation of around 1 to 2 percentage points. For the 75th percentile, the bias is more variable, with an overestimation of 2 to 3 percentage points in smaller samples that increases to 6 to 7 percentage points in larger samples. These findings highlight that even among generally well-performing methods, the precision of percentile estimates can vary meaningfully depending on sample size and the specific percentile being estimated. *Bootstrapping* maintains similarly low relative biases but has higher *Stability* values in very small samples, indicating less precision. The performance of *JK2* methods, particularly *JK2 Half* and *Full*, presents a somewhat perplexing pattern. While relative biases are generally higher compared to other estimators, the specific trends vary across percentiles and sample sizes. For the 25th percentile, relative biases are notably larger in smaller samples (ranging from 6 to 10 percentage points) and tend to decrease in larger samples (approximately 2 to 8 percentage points). In contrast, for the median, relative biases appear more consistent across both estimation methods and sample size groups. Interestingly, for the 75th percentile, this pattern is reversed: relative biases are relatively small in smaller samples (around 4 to 8 percentage points) but increase substantially in

[18] We also analyzed the percentiles 25th, 50th, and 75th by sub-group levels ($gender = girl$). The tables for these analyses are provided in the Appendix B 1, Appendix B 2, and Appendix B 3. The results overlap with the overall level, however *Relative Bias* is higher for variance estimators in the sub-group levels, as student sample size n is lower.

larger samples, reaching between 8 and 13 percentage points. These contrasting trends suggest that JK2's performance may be sensitive to the specific distributional characteristic being estimated, particularly in the context of non-smooth statistics. *Fay*-adjusted *BRR* methods perform better with lower adjustment factors (10% and 30%), while the 50% factor leads to higher *Relative Bias*, particularly in small samples. *Stability* for standard *BRR* is the best among all methods, underscoring its precision in this scenario.

In the ***Odd Number of Schools Scenario***, standard *BRR* and *Bootstrapping* continue to perform well, maintaining *Relative Bias*es below 5% across percentiles. However, *Bootstrapping* shows slightly higher *Stability* values than *BRR* in smaller samples. *Fay*-adjusted *BRR* methods perform acceptably at lower adjustments (10% and 30%) but exhibit more *Relative Bias* at the 50% level, with biases reaching 10 percentage points in small samples. *JK2* methods display consistent biases across percentiles but with higher values, often exceeding 7 percentage points for the 25th and 75th percentiles. *Stability* for *JK2* and its *Fay*-adjusted methods is notably worse compared to *BRR*, particularly in small sample sizes.

Under the ***Non-Response Scenario***, the increased complexity leads to reduced performance for most methods. Standard *BRR* and *Bootstrapping* again stand out, with *Relative Biases* generally below 3 percentage points in larger samples. *BRR* with *Fay* 50% *OrgStr* shows significant overestimation, with biases reaching as high as 16 percentage points in small samples. *JK2 OrgStr* methods, conversely, underestimate sampling variance, as indicated by *Relative Bias* values below 1 in small samples. *Stability* deteriorates further for *JK2* methods, especially for the 75th percentile in very small samples, where *Stability* values exceed 0.50. *Bootstrapping* remains a viable choice, particularly in medium sample sizes, balancing between *Relative Bias* and *Stability* more effectively than other methods.

In summary, standard *BRR* proves to be the most reliable variance estimator across scenarios, offering low *Relative Bias* and strong precision (low *Stability* values). *Bootstrapping* emerges as a close alternative, particularly for reducing *Relative Bias*, though it exhibits slightly worse *Stability* in small samples. *Fay*-adjusted *BRR* methods perform well at lower adjustment factors but are less effective with 50% adjustments in complex designs. *JK2* methods show greater *Relative Bias* and higher *Stability* values, especially for percentiles, highlighting the need for careful selection and calibration of estimators based on study design and objectives.

6.5 Benchmark Percentages (Low, High, and Advanced Benchmarks)

Relative Bias and *Stability* are presented for Low, High, and Advanced International Benchmarks in **Table 11**, **Table 12**, and **Table 13**.

[pls insert **Table 11**, **Table 12**, and **Table 13** about here]

For ***Regular Sample Scenario***, differences among variance estimators are minimal. Standard *BRR* demonstrates consistent performance, with *Relative Bias* values close to 1 across sample sizes and low *Stability* values, reflecting reliable precision. *Fay*-adjusted *BRR* variants show similar results, with *Relative Bias*es approximately at 1.01 and *Stability* values consistently low (e.g., 0.28 in very small samples). *Bootstrapping* also performs well, achieving slightly lower *Relative Bias*es (around 0.99) but marginally higher *Stability* in very small samples. Similarly, *JK2 Full* and *Half* methods exhibit comparable performance to *BRR*, with *Relative Bias* values at 1.01 and similarly low *Stability* values. These results confirm that all estimators perform well under regular sampling conditions, with only minor variations in *Stability*.

For the ***Odd Number of Schools Scenario***, results align closely with those from the ***Regular Sample Scenario***. *BRR*, its *Fay*-adjusted variants, and *JK2* methods maintain *Relative Bias* values close to 1, indicating minimal overestimation or underestimation. *Stability* values are slightly higher than in the regular scenario but remain low overall, highlighting good precision across methods. *BRR* with a 50% *Fay* adjustment performs slightly worse, with *Relative Bias* reaching 1.03 in very small samples and slightly higher *Stability* (e.g., 0.28). The application of Fay's method appears suboptimal in the context of BRR, as the relative bias increases with higher values of the Fay factor. This pattern is not observed in the JK2 estimator, which remains more stable across different Fay factor levels. Notably, the behavior of Fay's method within BRR differs for the advanced benchmark: it initially produces substantial underestimation when no Fay adjustment is applied, but the bias gradually diminishes and approaches zero as the Fay factor increases. Despite this, the differences remain modest, and *JK2* estimators often exhibit slightly better *Stability* performance than *BRR*. These findings suggest that odd-number scenarios do not introduce significant variation in estimator performance.

In the ***Non-Response Scenario***, the differences among estimators become more pronounced. *JK2* estimators, especially *OrgStr* variants, show lower *Relative Bias* performance compared to other scenarios, with values just below 1 in some cases, indicating

slight underestimation. *Stability* values for *JK2 Full* and *Half* remain relatively low (e.g., 0.28 for very small samples), contributing to their strong performance. *BRR* and *Fay*-adjusted *BRR* variants also perform well, with *Relative Bias* values near 1 and low *Stability* values (e.g., 0.28 in very small samples for standard *BRR*). However, *BRR OrgStr* tends to overestimate more notably, with *Relative Bias* values above 1 (e.g., 1.08 for very small samples), though its *Stability* remains acceptable. *Bootstrapping* exhibits reasonably good performance, with biases slightly below 1 and slightly higher *Stability* than *BRR* and *JK2* methods, particularly in smaller samples.

In summary, while the regular and odd-number scenarios show negligible differences in estimator performance, the non-response scenario reveals more substantial distinctions. *JK2* estimators emerge as strong performers in terms of both *Relative Bias* and *Stability*, while *BRR* and its *Fay* variants remain dependable. The BRR and JK2 methods behave similarly across all benchmark levels. For the advanced benchmark, both methods severely underestimate relative bias in medium and large sample scenarios. Interestingly, for the same benchmark, both methods demonstrate unexpectedly high stability in very small and small samples, given the extremely low number of students in those cases. Differences are most notable under complex sampling conditions, highlighting the importance of careful estimator selection in these contexts.

6.6 Correlations

Table 14 indicates *Relative Bias* and *Stability* for the Pearson correlation estimate between score 1 and score 2.

[pls insert **Table 14** about here]

In ***Regular Sample Scenario***, *BRR* and its *Fay* variants deliver commendable accuracy, with *Relative Bias* values close to 1.00 for small, medium, and large samples. *Stability* improves with increasing sample size, as evidenced by *BRR Fay* 50% achieving a *Stability* value of 0.12 in large samples. *Fay*-adjusted variants show slightly reduced *Stability* at smaller sizes (e.g., *BRR Fay* 10% at 0.25 for very small samples), but this difference diminishes with larger samples, where all *BRR* variants converge towards excellent precision. *JK2* methods show comparable *Relative Bias* to *BRR*, occasionally performing marginally better. Notably, *JK2* demonstrates slightly improved stability in smaller sample sizes, suggesting it may offer a minor advantage in situations with limited data, particularly when estimating non-smooth statistics.

For ***Odd Number of Schools Scenario***, *BRR* and its *Fay* variants maintain their slightly better performance in *Relative Bias* but weaker performance in *Stability*. The standard *BRR*, as well as *Fay*-adjusted versions, show minimal *Relative Bias* across sample sizes, except very small samples where the *standard BRR* and its *10%* and *30%* variants underestimate by 6 to 7 percentage points. In *Relative Bias*, *BRR with 50% Fay* is better across different sample size scenarios and shows marginally increased variability in small sizes. *Stability* remains competitive, with *BRR Fay* 10% and 30% exhibiting better precision (e.g., *Stability* values around 0.20 for small samples). Fay's method has no impact on JK2 performance, and JK2 exhibits relatively weaker performance in *Relative Bias* under very small sample sizes. Nonetheless, *Stability* for JK2 remains comparable to BRR in most cases.

In the ***Non-Response Scenario***, *BRR OrgStr* shows relatively strong performance in medium and large sample sizes. It achieves *Relative Bias* values close to 1.00 and *Stability* measures as low as 0.15–0.17. However, *Stability* remains volatile across all BRR variants, especially in smaller sample sizes. In those smaller samples, *BRR* with lower *Fay* adjustments (*10%* and *30%*) performs better in terms of *Relative Bias*. The *standard BRR* and *Fay 50%* variants tend to slightly overestimate in smaller samples, but *Stability* improves as sample sizes increase. *Bootstrapping* consistently underestimates sampling variance, with *Relative Bias* (e.g., 0.83–0.98). *JK2 Full* and *Half* variants perform reasonably well in very small and small samples but exhibit similar underestimation patterns to BRR in certain conditions. Notably, both *JK2 OrgStr* variants substantially underestimate sampling variance across all sample sizes, with *Relative Bias* values falling to as low as 0.58–0.65 in small and medium samples. Despite having relatively low stability values (e.g., 0.08–0.19).

6.7 Summary across all Results

For the purpose of summarizing overall performance, we compute the average across all variance estimators for the statistics that are reported in this study. Specifically, this includes: (i) mean estimates when grouping is overall and girls; (ii) difference in mean estimates between girls and boys, (iii) benchmark percentages for low, high, and advanced benchmarks, when grouping is overall; (iv) Pearson correlation estimates between Score 1 and Score 2; and (v) percentile estimates (25th, median, and 75th) when grouping is overall and for girls. In total, this corresponds to three mean statistics, along with key benchmarks

and distributional indicators. The aim of this summary is to provide a high-level overview of the overall behavior and comparative performance of all variance estimators across commonly reported outcomes in ILSAs. We acknowledge the limits of such approach, such as equal contribution of all scenarios/analysis types to the average.

Table 15 compares variance estimators across a variety of smooth (e.g., means, correlations) and non-smooth (e.g., benchmarks, percentiles) statistics, providing a comprehensive evaluation of their performance under different estimation conditions. **Figure 1**, **Figure 2**, and **Figure 3** visually summarize these results for the ***Regular Sample***, ***Odd Number of Schools***, and ***Non-Response* s**cenarios, analyzing both *Relative Bias* and *Stability* across sample sizes.

[pls insert **Table 15**, **Figure 1**, **Figure 2**, and **Figure 3** about here]

In the figures, *Relative Bias* is depicted on the y-axis, with a value of 1 (shown by the red line) indicating no bias, values above 1 reflecting overestimation, and values below 1 indicating underestimation, by the spread of the boxplots. *Stability* is illustrated by the error bars, with narrower bars representing greater precision.

6.7.1 Regular Sample Scenario

In the ***Regular Sample Scenario***, according to **Table 15** and **Figure 1**, Bootstrapping emerges as the top performer in terms of *Relative Bias*, consistently achieving values close to 1.00 across all sample sizes. This is followed closely by *BRR* and its *Fay* variants, which maintain slightly higher *Relative Bias* but still demonstrate good accuracy. *JK2 Full* and *JK2 Half* exhibit slightly higher *Relative Bias* compared to *BRR*, though the two *JK2* estimators are nearly identical in performance. *Stability*-wise, *BRR* outperforms the *JK2* and *Bootstrapping* estimators, with lower variability, especially in smaller sample sizes. *Fay* adjustments for *JK2* lead to further worsening of *Relative Bias* and *Stability*, particularly for the *Fay* 50% variant. *Stability* improves for all estimators as the sample size increases, but *BRR* slightly demonstrates better performance compared to *JK2* estimators. Interestingly, *JK2* exhibits smaller *Relative Bias* for smooth statistics but larger *Relative Bias* for non-smooth statistics, as reflected in the broader margins of its boxplots.

For all replication methods, the precision of variance estimators improves with increasing sample sizes, as evidenced by lower *Stability* values. *BRR and BRR with 10% Fay* remains the most consistent across all sample sizes in *Stability*, demonstrating better performance compared to the *JK2* estimators.

6.7.2 Odd Number of Schools Scenario

For the ***Odd Number of Schools Scenario***, *BRR* and its *Fay* variants (10% and 30%) retain their strong performance, with *Relative Bias* near the ideal level across all sample sizes, as indicated in **Table 15** and **Figure 2**. However, the *Fay* 50% variant underperforms compared to the 10% and 30% variants, with larger variability in smaller sample sizes. *Bootstrapping* also performs well, particularly in larger sample sizes, where it closely matches the performance of *BRR*. *Stability* remains a strong point for *BRR* and *Bootstrapping*, with error bars generally narrower than those observed for *JK2* estimators.

The *JK2* methods outperform *BRR* and *Bootstrapping* in very small samples in terms of *Relative Bias*. Among *JK2* methods, the *Half* estimator slightly outperforms the *Full* estimator in terms of both *Relative Bias*, while the *Full* estimator is better in terms of *Stability* than the *Half* estimator. However, the *Fay* adjustments for *JK2* again degrade performance, with the *50*% variant showing the largest deterioration.

6.7.3 *Non-Response Scenario*

As indicated in **Table 15** and **Figure 3** in the ***Non-Response Scenario***, *BRR* and its *Fay* variants continue to perform robustly, however only in larger sample sizes, where *Relative Bias* is closer to 1.0 and *Stability* is improved. In medium sample sizes, *JK2 Half* and *Full* estimators perform the best and follow closely *BRR* in large sample sizes. In smaller sample sizes, *Bootstrapping* exhibits better performance in terms of *Relative Bias*, but its *Stability* does not match that of *BRR*. The *BRR Fay* 50% *OrgStr* variant significantly overestimates sampling error, while *JK2 OrgStr* variants considerably underestimate sampling error.

Stability remains a key distinction, with *JK2 OrgStr* variants showing better *Stability* in smaller sample sizes compared to other *JK2* methods. However, overall, *BRR*, *Bootstrapping, and JK2* perform similarly in terms of *Relative Bias*, however *JK2* performs worse in *Stability,* compared to *BRR* and *Bootstrapping*.

6.7.4 Summary

Across all scenarios, the following patterns are evident:

1. Across all panels presented in **Figure 1**, **Figure 2** and **Figure 3,** corresponding to increasing sample sizes from very small (30 schools, ~900 students) to large (150

schools, ~4500 students), the *Stability* indicator of the estimators decreases, indicating improved precision with larger sample sizes. This difference is especially evident when the sample sizes are compared between 50 and 100 schools.

2. *BRR* and its *Fay* variants are consistently strong performers, particularly for large samples. *Stability* remains superior to *JK2* estimators across all scenarios, although *Fay* adjustments (especially 50%) lead to minor increases in *Relative Bias* and variability.
3. *Bootstrapping* performs exceptionally well in smaller samples in both regular and non-response scenarios, with low *Relative Bias* and competitive *Stability*. However, its performance slightly dips in the *odd-number-of-schools scenarios*.
4. *JK2* estimators perform well closely following *BRR* and *Bootstrapping* in both *Relative Bias*. However, their precision is more volatile in terms of *Stability,* especially in smaller sample sizes. On the other hand, the *JK2* (*Full* and *Half)* outperforms *BRR with 50%* in terms of *Relative Bias*, especially in *Odd Number of Schools* scenario. *Fay* adjustments for *JK2*, particularly 50%, exacerbate these shortcomings. However, *JK2 OrgStr* variants demonstrate the best *Stability* in non-response scenarios, albeit with significant underestimation of *Relative Bias*.

Overall, these findings suggest that *BRR*, *JK2*, and *Bootstrapping* methods work well under common conditions as reflected in the ***Regular Sample Scenario***, but greater caution is required when there are uneven numbers of schools in a variance stratum (***Odd Number of Schools Scenario)*** and PSU non-response (***Non-Response Scenario)***. In the ***Odd Number of Schools Scenario***, *BRR* with *Fay* 50% tends to overestimate variance, while *JK2 Full* and *Half* remain stable and exhibit minimal differences between their *Fay*-adjusted variants (only 10% and 30% Fay). Keeping non-responding PSUs in a variance stratum (*BRR OrgStr*) or pairing units within PSUs (*JK2 OrgStr*) introduce *Relative Bias—BRR OrgStr* through overestimation and *JK2 OrgStr* through underestimation of sampling errors, making them less reliable under PSU non-response. *Bootstrapping* remains a strong alternative across both smooth and non-smooth statistics, providing stable estimates across different conditions.

7 Discussion and Recommendations

For practical applications in ILSAs, it is important to identify a single variance estimation method that is both accurate and precise under common sampling scenarios and

for different types of population parameter estimators, as employing different methods for varying statistics is logistically challenging. With this reason, this comprehensive report highlights the importance of examining and discussing the merits and strengths of the spectrum of variance estimators found in the field. However, as the results show, there are distinct advantages and disadvantages associated with each method. Our findings, in line with Judkins (1990) and Valliant et al. (2018), confirm that replication methods such as BRR and *JK2* perform well in larger sample sizes without extreme outliers, demonstrating low *Stability* and low *Relative Bias*. This supports their use for variance estimation in large-domain analyses typical of ILSAs. However, in smaller samples or subgroups determined by clusters, variance estimates become less precise, which is evident in the ***Regular Sample Scenario***.

7.1 Is *BRR* an alternative in all contexts?

Interpreting our findings, one could recommend adopting *standard BRR* as the primary method due to its accuracy in terms of *Relative Bias* and precision for large-domain estimation. However, the issue of missing replicates, caused by the omission of half-samples in each sample, makes it unsuitable for smaller-domain estimation, such as estimating sampling errors for the achievement of girls by socioeconomic status where small n can be observed in a few PSUs. If all possible observations are omitted according to the *BRREP* and *Hadamard* matrix in a given replicate, the SE cannot be computed in a mathematical way. That being said, *Standard BRR* still performs relatively well in small sample sizes (e.g., 30 schools), as demonstrated in our results, when compared to other Fay variants. However, in small sample size scenarios, particularly for sub-group analyses, *BRR* is outperformed by *JK2* in terms of *Stability*.

This highlights that while *BRR* performs well overall, it encounters a critical mathematical limitation in small-domain estimation due to missing replicates. Robert *Fay* (1984) addressed this issue by introducing the *Fay* factor, which modifies the weights to overcome this limitation. The natural question that follows is: What *Fay* factor should be used to best mirror the performance of *Standard BRR*?

7.2 What *Fay* factor to use in *BRR*?

Judkins' suggestion of using "*gentler*" *Fay* factors instead of *BRR* with 50% aligns with our findings, as we observe that *Fay* factors of 10% and 30% outperform a *Fay* factor

of 50%. A 10% *Fay* factor closely mirrors standard *BRR*'s performance, especially in odd-numbered PSU scenarios, where higher *Fay* factors lead to overestimation. It is not surprising that *standard BRR* outperforms *BRR with 50% Fay* in many scenarios, as Judkins (1990) identified *Fay*'s method as a compromise between *BRR* and *JK2*, intended to balance the strengths of both for smooth and non-smooth statistics.

Our findings suggest prioritizing *standard BRR* for most scenarios, with the 10% *Fay* adjustment used selectively where a smoother "*compromise*" is required. The similarity between the 10% *Fay* factor and standard *BRR* may stem from the 10% factor's closer resemblance to a 0% *Fay* adjustment, a hypothesis warranting further investigation. However, for small-domain estimates typical in ILSAs—such as subgroup achievements by socio-economic status—*BRR* remains unsuitable due to its mechanical problem with missing replicates. With this reason, we recommend *BRR with Fay 10%* as a viable alternative, as it addresses the challenges inherent in *Standard BRR* while maintaining reliable performance. This adjustment could serve as an alternative for ILSAs currently using *BRR with Fay 50%* (e.g., TALIS and PISA), particularly when targeting small-domain estimation (See TALIS 2018 International Report [OECD, 2019]). Given the advantages demonstrated by *BRR with Fay 10%*, study centers are encouraged to further investigate its suitability and determine whether it could effectively replace the existing method in certain applications.

7.3 What *Fay* factor to use in *JK2*?

Fay factors introduced to the *JK2* method perform identically for smooth statistics, offering a potential alternative when *JK2* cannot mathematically compute due to very small sample sizes or orphan cases where *JKREP* is assigned a value of 0, such as in school-level analyses. Introducing a Fay factor to JK2 could help produce sampling error estimates in such cases; however, this does not necessarily mean the resulting estimates are more accurate or reliable. Variance estimation for sub-groups with small sample sizes remains inherently problematic, and this issue warrants further research.

Another motivation for introducing Fay factors into *JK2* was to enhance its performance for non-smooth statistics such as percentiles. However, our results indicate that *Fay* factors do not improve *JK2*'s performance in this context; in fact, higher *Fay* factors—particularly the *50%* level—tend to worsen performance. In contrast, the *10%* Fay factor yields results more comparable to the *JK2 Half method without Fay*,

outperforming the *30%* and *50%* variants. Despite these findings, we recommend further investigation into the theoretical inclusion of *Fay* factors, particularly for deriving alternatives for percentile estimation. For instance, theoretical adaptations like those suggested by Woodruf (1952), supported by Judkins (1990), could offer potential pathways for improving *JK2*'s applicability to non-smooth statistics.

7.4 *JK2 Half* vs *JK2 Full*: is it worth running twice the number of replicates?

JK2 Half emerges as a strong alternative to *JK2 Full* due to its comparable performance—at least within visible and relevant accuracy and precision margins—and significantly reduced computational burden. In the context of ILSAs, especially where PVs are used, the doubled runtime required by *JK2 Full* becomes a notable disadvantage, particularly when the number of PVs range from five to ten or more. For instance, consider the TIMSS 2023 study involving 60 countries, each with approximately 100 variance strata, and 5 PVs. Using the *Half* method, one would compute $100 \times 5 \times 60$ replicates for all countries for a single statistic with overall results. In contrast, the *Full* method requires $100 \times 5 \times 60 \times 2$, doubling the computational effort. Previous literature also questions the need for two replicates per stratum. Judkins (1990, p. 229) suggests that there is little reason to investigate the full *Jackknife*, and Valliant et al. (2018, p. 446) note that deriving two replicates per variance stratum may only increase the computational burden, especially when the number of variance strata is large. Our findings support this view, showing that *JK2 Half* maintains accuracy and precision with far less computation.

In addition, our results indicate that the *Half* method very slightly outperforms the *Full* method in *Relative Bias* (see the comparison of the *JK2 Full* and *Half* in **Table 15**), particularly when between-PSU variance is ignored within a given variance stratum due to scenarios such as facing an odd number of PSUs or non-response. This further underscores the practicality of *JK2 Half* for ILSAs, especially when computational efficiency and accurate performance are critical. Given these findings, we encourage study centers—particularly those responsible for methodological frameworks in assessments like TIMSS and PIRLS—to evaluate the continued use of *JK2 Full* in light of these considerations.

While *JK2 Full* demonstrates marginally better stability for percentiles and correlations in smaller samples, this improvement is relatively minor and may not justify the significantly higher computational cost, especially in studies with larger sample sizes. That said, if a study places a high priority on stability for these specific statistics in small-

sample contexts, *JK2 Full* remains a valid—albeit more resource-intensive—option. Ultimately, the choice between these two methods should reflect the study's specific priorities: whether that is maximizing computational efficiency or achieving slightly better stability in variance estimation. For instance, if a study center aims to improve the precision of imputation error by increasing the number of PVs (e.g., from 5 to 10), *JK2 Half* presents a reliable and efficient alternative. However, users should be aware that neither method resolves the limitations *JK2* has with percentiles, which may require further methodological development.

7.5 Is *Bootstrapping* the alternative variance estimator?

Bootstrapping provides an effective alternative for variance estimation, particularly excelling in quantile estimation, and outperforming methods like *JK2* in this specific context. Its computational simplicity, relying on variance strata and PSU indicators, makes it an accessible method for researchers. In this study (see footnote 14), we demonstrated how *Bootstrapping* can be easily applied to a dataset using a variance stratum indicator (e.g., *JKZONE*) and a PSU indicator (e.g., *IDSCHOOL*). However, *Bootstrapping* faces a significant limitation: its inherent variability across replicate sets. This variability results in slight changes in SE estimates with each *Bootstrapping* run unless a random seed is specified. While this deviation highlights the uncertainty associated with SE estimates, it may confuse users trying to reproduce tables in published reports.

To address this limitation, researchers should consider *Bootstrapping* as an alternative replication method, ensuring reproducibility by specifying random seeds. For international reports requiring fixed estimates, study centers could explore its application further and consider its use selectively for specific analyses, such as quantile estimation, where its strengths are most evident.

7.6 How should variance strata be constructed when *Odd Numbers of PSUs* or *Non-Response* occurs?

Variance strata and PSU pairing require special consideration to adequately simulate the variability in the estimation induced by random sampling design, while preventing orphan cases, particularly in scenarios involving ***Odd Numbers of PSUs*** or ***Non-Response***. As Valliant et al. (2018) noted, pairing of PSUs through collapsing variance strata to reduce the number of replicates often leads to an overestimation of

sampling error, while constructing a variance stratum with a single PSU leads to underestimation, especially for the *Jackknife*. In the ***Odd Number of PSUs*** scenario, *BRR* addresses this by collapsing two variance strata into one, forming triplets of PSUs within a single variance stratum. Our results confirm that this approach, particularly with *BRR with 50% Fay*, tends to overestimate the sampling error. However, this overestimation is mitigated by *Standard BRR* or *BRR* with lower *Fay* factors such as 10%. In these cases, *JK2* handles uneven PSU numbers by pairing sub-PSUs (e.g., students or classes), thereby ignoring between-PSU variance in a given replicate. As this adjustment typically affects only a single variance stratum, the resulting underestimation has minimal impact on overall sampling error, as evidenced by our findings. That said, a potential enhancement to *JK2* would be the development of a true triplet structure to more appropriately account for orphaned PSUs.

In ***Non-Response*** scenarios, the issue becomes more pronounced. *JK2 OrgStr* treats multiple variance strata with single PSUs by pairing sub-PSUs (e.g., students or classes) as if they were full PSUs, which leads to significant underestimation of the sampling error. Valliant et al. (2018) illustrated this issue with the SMHO hospital survey, where pairing hospital beds (within-PSU) underestimated variance compared to pairing hospitals (between-PSU). Similarly, in ILSAs, pairing students within the same school risks underestimating sampling variance compared to pairing schools, particularly under *JK2 OrgStr*. On the other hand, for single-PSU variance strata under ***Non-Response*** scenarios, *BRR OrgStr* rotates the inflating or deflating contribution of this single-PSU across the Hadamard matrix. *BRR OrgStr* handles this in a way that overestimates the sampling error, contrary to *JK2 OrgStr*, which tends to underestimate it. This finding on *BRR OrgStr* has not been directly addressed in previous literature. One possible explanation could be the systematic inclusion of replicate weights, where the sum of weights introduced in *BRR OrgStr* may be higher than in other *BRR* variants. However, this hypothesis warrants further investigation to better understand the mechanism and its impact on variance estimation.

We believe that much research is needed to understand the theoretical and practical implications of non-response on the use of re-sampling variance estimators such as the *Jackknife* and the *Balanced Repeated Replication*. In this research, we have discussed at large that ILSAs often face a choice about how to pair schools into variance zones when a data collection is subject to school non-participation. That is, whether variance zones should strictly reflect the sampling design guiding a data collection, which would lead to

variance zones where a participating school is paired with a non-participating school; or whether they should reflect the structure of the data collected, in which case only participating schools would be paired.

Theoretically, the variance of an estimated parameter is an inferential problem arising from the finite sample paradigm. That is, everything else constant, non-response reduces the amount of information available for estimation, thereby widening the sampling distribution of the parameter of interest. To this end, the problem of non-response can be studied by examining the performance of variance estimators under different sample sizes. However, non-response also creates other inferential problems that do not disappear as sample sizes increases without limit, such as the problem we face when deciding whether to create variance strata before or after a data collection. These problems require an investigation that goes beyond comparing different sampling distributions, and we believe this is an interesting topic for future research.

7.7 Summary

In conclusion, our findings suggest that *BRR* and *JK2* are practical and robust methods for variance estimation in ILSAs, with *Bootstrapping* offering a viable alternative for secondary analyses or specific statistical needs. These methods should be carefully tailored to the specific design and sample characteristics, with attention to variance strata, PSU pairing, and the unique challenges posed by ***Odd Number of PSUs*** and ***Non-Response*** scenarios. By addressing these considerations, researchers can achieve both accuracy and precision in variance estimation while maintaining computational efficiency. Further investigation into refinements for *Fay* factors, *JK2* adjustments, and *Bootstrapping* reproducibility will be critical to advancing variance estimation methodologies in large-scale assessments.

Institutions such as *IEA*[19] providing datasets and analysis tools like the *IDB Analyzer* may consider accommodating multiple SE estimation methods for secondary research by including the necessary variables (e.g., replicate weights for *JK2* and *BRR*) to let scholars choose the most suitable estimation method, depending on their research questions and sample conditions. Supporting researchers in this process, they may

[19] www.iea.nl

consider developing an algorithm that selects the most suitable estimation method based on sample size and analysis type. Additionally, datasets and tools should allow researchers to utilize *Bootstrapping* for enhanced flexibility in variance estimation.

However, it is essential to recognize the limitations of these proposals. Implementing such changes can pose significant challenges in terms of cost and resource allocation, as developing these tools and creating user-friendly interfaces requires substantial investment. Moreover, transitioning to a new method or providing multiple options may confuse non-technical users and complicate communication with lay audiences. Since ILSA results are often used for high-stakes policy decisions, clarity and transparency in methodological changes are critical to maintaining trust and credibility. These challenges underscore the importance of carefully balancing methodological advancements with effective communication strategies to ensure that end users and policymakers fully understand the implications of any proposed changes.

8 Conclusions and Limitations

This study makes an important contribution to the understanding of sampling variance estimators in the context of ILSAs. Identifying optimal variance estimation techniques is essential for ensuring the precision and reliability of statistical inferences drawn from these large-scale assessments. With this study, we enhance the ILSA survey methodology by providing empirical insights into the precision of different variance estimators across various sampling conditions. For instance, if a variance estimator underestimates sampling error, it could result in a higher t-value than warranted in a significance test between two subgroups. Conversely, if a variance estimator significantly overestimates sampling error, the t-value could be understated, leading researchers to conclude there is no difference between subgroups when there is actually a difference present in the population. Such misestimations are not limited to subgroup comparisons but extend to other statistical comparisons, such as trend analyses within a country over multiple years.

A key contribution of this study is its comprehensive evaluation of variance estimation methods, examining *Fay* factors, the treatment of PSUs in non-response scenarios, and *Bootstrapping* applications for clustered sampling. By systematically comparing these methods across a range of scenarios, the research provides valuable insights to enhance ILSA survey methodology. Importantly, the study finds that replication methods, the current

standard in ILSAs, perform well even in small sample sizes of 30 PSUs, reinforcing their suitability for these assessments.

As in any research project, this study is not without limitations. Its methodology—based on an extensive simulation setup—also constrains its applicability to real-world data complexities. Real-world ILSA data often involve varying degrees of between-PSU and within-PSU variation, more intricate stratification structures with explicit and implicit strata, and significantly different population sizes. Additionally, real-world non-response is more complex, potentially occurring across multiple levels, which was not fully simulated here. Sample sizes also vary substantially across countries, adding another layer of complexity not captured in this study. Furthermore, the research focuses primarily on a single simulation setup with specific scenarios, limiting its generalizability. Another limitation is the scope of statistical parameters considered. While the study extensively investigates variance estimators, it does not include analyses like linear regression or scenarios with very small sub-group sizes, which could further challenge variance estimation methods.

Future research should expand on these limitations, incorporating more complex real-world scenarios, varied sample sizes, and additional statistical parameters. While this study provides a robust foundation and valuable guidance, continued examination of variance estimator performance under diverse conditions is essential. This research aims to serve as a cornerstone for further exploration, offering insights and benchmarks for future studies to refine and improve variance estimation in ILSAs.

9 References

Atasever, U., Huang, F.L. & Rutkowski, L. (2025). Reassessing weights in large-scale assessments and multilevel models. *Large-scale Assessments in Education,* 13, 9. https://doi.org/10.1186/s40536-025-00245-y

Atasever, U., Jerrim, J., & Tieck, S. (2024). Exclusion rates from international large-scale assessments: An analysis of 20 years of IEA data. *Educational Assessment, Evaluation and Accountability, 36*(3), 405–428. https://doi.org/10.1007/s11092-023-09416-3

Austin, P. C. (2024). A comparison of variance estimators for logistic regression models estimated using Generalized Estimating Equations (GEE) in the context of observational health services research. *Statistics in Medicine*, *43*(29), 5548–5561. https://doi.org/10.1002/sim.10260

Beaumont, J.-F., & Émond, N. (2022). A bootstrap variance estimation method for multistage sampling and two-phase sampling when Poisson sampling is used at the second phase. *Stats, 5*(2), 339–357. https://doi.org/10.3390/stats5020019

Bruch, C., Munnich, R., & Zins, S. (2011). Variance Estimation for Complex Surveys. *Advanced Methodology for European Laeken Indicators.*

Choudhry, G. H., & Valliant, R. (2002). WESVAR: Software for Complex Survey Data Analysis.

Cortes, D., Hastedt, D. & Meinck, S. (2025). Evaluating uncertainty: the impact of the sampling and assessment design on statistical inference in the context of ILSA. *Large-scale Assessments in Education* 13, 10. https://doi.org/10.1186/s40536-025-00246-x

Demnati, A., & Rao, J. N. K. (2002). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, December 2010, Vol. 36, no. 2.

Dippo, C. S., Fay, R. E., & Morganstein, D. H. (1984). Computing variances from complex samples with replicate weights. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 489–494). American Statistical Association.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, December 1999, Vol. 25, no. 2.

Dumais, J. Meinck, S. (2013). Estimation of weights, participation rates, and sampling error. In M.T. Tatto (Ed.), *TEDS-M: Technical report* (pp.129–160). International Association for the Evaluation of Educational Achievement (IEA). https://www.iea.nl/sites/default/files/2019-05/TEDS-M_technical_report.pdf

Efron, B. (1982). The jackknife, the *bootstrap*, and other resampling plans. *Society for Industrial and Applied Mathematics* (SIAM). https://doi.org/10.1137/1.9781611970319

Fay, R. E. (1984). Some properties of estimates of variance based on replication methods. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 495 –500). American Statistical Association.

Fishbein, B., Siegel, P., Atasever, U., & Cooney, D., & Gonzalez, G. (2024). Standard error estimation in TIMSS. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 technical report (methods and procedures)* (pp. 13.1–13.10). TIMSS & PIRLS International Study Center, Boston College. https://doi.org/10.6017/lse.tpisc.timss.rs6950

Foy, P. & Joncas, M. (2004). TIMSS 2003 sampling design. In M. O. Martin, I. V. S. Mullis & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 108–123). TIMSS & PIRLS international study center, Boston College.

Foy, P. (2005). Estimating and interpreting variance components in international comparative studies in education. *Studies in Educational Evaluation, 31*(2–3), 173–191. https://doi.org/10.1016/j.stueduc.2005.05.009

Gonzalez, E. J. (1996). Standardizing the TIMSS international scale scores. In M. O. Martin, D. L. Kelly (Eds.), *TIMSS 1995 technical report.* Boston College.

Gonzalez, E. J., & Foy, P. (1996). Estimation of sampling variance. In M. O. Martin, D. L. Kelly (Eds.), *TIMSS 1995 technical report*. Boston College.

Gonzalez, E. J., & Foy, P. (2000). Estimation of sampling variance. In M. O. Martin, K. D. Gregory, & S. E. Semler (Eds.), *TIMSS 1999 technical report*. Boston College.

Gonzalez, E. J., & Kennedy A. M. (2003). Statistical analysis and reporting of the PIRLS data. In M. O. Martin, I. V.S. Mullis, A. M. Kennedy (Eds.), *PIRLS 2001 technical report*. Boston College.

Gonzalez, E. (2024). Calculating standard errors in ILSAs. Document presented at the TIMSS 2023 International Database (IDB) Workshop, July 2024.

Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education, 9*(1), 17. https://doi.org/10.1186/s40536-021-00109-1

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.

IEA. (2025). Help Manual for the IEA IDB Analyzer (Version 5.0). (Available from www.iea.nl)

Johnson E. G., & Rust K. F. (1992a). Effective df for variance estimates from a complex sample survey. In Proceedings of the section on survey research methods. *American Statistical Association*. (Preprint copy obtained from K. Rust, September 2024).

Judkins, D. (1990). Fay's method for variance estimation. *Journal of Statistics Sweden*, Vol 6., No.3, 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* (pp. 141–159). Springer.. https://doi.org/10.1007/978-981-16-9142-3_8

Keslair, F. (2020). Analysing PIAAC data with stata. In D. B. Maehler & B. Rammstedt (Eds.), *Large-scale cognitive assessment* (pp. 149–164). Springer International Publishing. https://doi.org/10.1007/978-3-030-47515-4_7

Kish, L. (1965). *Survey sampling*. Wiley.

Kovar, J. G. (1985). *Variance Estimation of Nonlinear Statistics in Stratified Samples*. Methodology Branch Working Paper No. 85-052E. Statistics Canada.

Kovar, J. G., Rao, J. N. K., & Wu, C. F. J. (1988). Bootstrapping and other methods to measure errors in survey estimates. *Canadian Journal of Statistics, 16*(S1), 25–45. https://doi.org/10.2307/3315214

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. https://doi.org/10.1214/aos/1176345580

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

Mahdizadeh, M., & Zamanzade, E. (2021). New estimator for the variances of strata in ranked set sampling. *Soft Computing, 25*(13), 8007–8013. https://doi.org/10.1007/s00500-021-05787-1

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), 6. https://doi.org/10.1186/s40536-021-00099-0

Mang, J., Küchenhoff, H., & Meinck, S. (2024). Evaluating German PISA stratification designs: A simulation study. *Large-Scale Assessments in Education, 12*(1), 15. https://doi.org/10.1186/s40536-024-00203-0

McCarthy, P. J. (1969): Pseudo-replication: Half samples. *Review of the International Statistical Institute, 37*(3), pp. 239–264.

Meinck, S. (2020). Sampling, weighting, and variance estimation. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment* (Vol. 10, 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: Methods and implications. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education* (pp. 1–25). Springer International Publishing. https://doi.org/10.1007/978-3-030-38298-8_25-1

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.

Mirazchiyski, P. V. (2021). RALSA: The R analyzer for large-scale assessments. *Large-Scale Assessments in Education, 9*(1), 21. https://doi.org/10.1186/s40536-021-00114-4

OECD (2012), *PISA 2009 technical report.*. OECD Publishing. https://doi.org/10.1787/9789264167872-en

OECD. (2012). *Better skills, better jobs, better lives: A strategic approach to skills policies*. OECD Publishing. https://doi.org/10.1787/9789264177338-en

OECD (2019), *TALIS Starting Strong 2018 technical report.. OECD Publishing*. https://doi.org/10.1787/0921466e-en

OECD. (2019). *TALIS 2018 results (Volume I): Teachers and school leaders as lifelong learners*. OECD Publishing. https://doi.org/10.1787/1d0bc92a-en

OECD. (2020). TALIS 2018 results (Volume II): Teachers and school Leaders as valued professionals. OECD Publishing. https://doi.org/10.1787/19cf08df-en

OECD. (2024). PISA 2022 technical report. OECD Publishing. https://doi.org/10.1787/01820d6d-en

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

Rao, J. N., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, *83*(401), 231–241. https://doi.org/10.1080/01621459.1988.10478591

Rasch, B., Hofmann, W., Friese, M., & Naumann, E. (2010). Der t-Test. In B. Rasch, W. Hofmann, M. Friese, & E. Naumann, Quantitative Methoden (pp. 43–117). *Springer*. https://doi.org/10.1007/978-3-642-05272-9_3

Rust, K., & Rao, J. (1996). Variance estimation for complex surveys using replication techniques. Statistical *Methods in Medical Research, 5*(3), 283–310. https://doi.org/10.1177/096228029600500305

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 117–130). Taylor & Francis. https://doi.org/10.1201/b16061-11

Rutkowski, L., Davier, M. von, & Rutkowski, D. (Eds.). (2014). Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis. *CRC Press*. https://doi.org/10.1201/b16061

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

Sandoval-Hernandez, A., & Carrasco, D. (2020). Analysing PIAAC data with the IDB Analyzer (SPSS and SAS). In D. B. Maehler & B. Rammstedt (Eds.), *Large-scale cognitive assessment* (pp. 117–148). *Springer International Publishing*. https://doi.org/10.1007/978-3-030-47515-4_6

Schneider B. (2023). "svrep: Tools for Creating, Updating, and Analyzing Survey Replicate Weights". R package version 0.6.0

Siegel, P., Cooney, D., Ridenhour, J., & Atasever, U. (2024). TIMSS 2023 sampling implementation. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), TIMSS 2023 technical report (methods and procedures) (pp. 9.1-9.124). TIMSS & PIRLS International Study Center, Boston College. https://doi.org/10.6017/lse.tpisc.timss.rs7763

Siegel, P., & Foy, P. (2024). TIMSS sample design. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), TIMSS 2023 technical report (methods and procedures) (pp. 3.1-3.30). TIMSS & PIRLS International Study Center, Boston College. https://doi.org/10.6017/lse.tpisc.timss.rs7952

Snijders, T. A. B., & Bosker, R. J. (2011). Multilevel analysis: An introduction to basic and advanced multilevel modeling. *SAGE*.

Schulz, W. (2011). The reporting of ICCS results. In W. Schulz, J. Ainley, J., & J. Fraillon (Eds.), *ICCS 2009 technical report*. International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W. (2015). The reporting of ICILS results. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *ICILS 2013 technical report*. International Association for the Evaluation of Educational Achievement (IEA).

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society: Series B (Methodological), 11*(1), 68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika, 43*(3/4), 353–360.

Valliant, R. & Rust, K. (2010). Degrees of freedom approximations and rules-of-thumb. *Journal of Official Statistics, 26*(4), 585–602.

Valliant, R., Dever, J. A., & Kreuter, F. (2018). Practical tools for designing and weighting survey samples. *Springer International Publishing*. https://doi.org/10.1007/978-3-319-93632-1

Von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., Bookbinder, A., Bezirhan, U., & Yin, L. (2024). *TIMSS 2023 international results in mathematics and science*. TIMSS & PIRLS International Study Center, Boston College. https://doi.org/10.6017/lse.tpisc.timss.rs6460

Tukey, J. W. (1958): Bias and confidence in not quite large samples (Abstract). *Annals of Mathematical Statistics*, 29, p. 614.

Wolter, K. M. (2007). Introduction. In *Introduction to variance estimation* (pp. 1–20). *Springer*. https://doi.org/10.1007/978-0-387-35099-8_1

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, *47*(260), 635–646. https://doi.org/10.1080/01621459.1952.10483443

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27. https://doi.org/10.1111/j.1745-3992.2010.00190.x

Zhang, T., Bailey, P., Liao, Y., & Sikali, E. (2024). EdSurvey: An R package to analyze large-scale educational assessments data from NCES. *Large-Scale Assessments in Education, 12*(1), 41. https://doi.org/10.1186/s40536-024-00222-x

10 Tables

Table 5: Variance Estimator Comparison - Means Statistics (Overall)

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.02	1.01	1.01	1.03	0.20	0.15	0.10	0.08
BRR with Fay 10%	1.02	1.01	1.01	1.03	0.20	0.15	0.10	0.08
BRR with Fay 30%	1.02	1.01	1.01	1.03	0.20	0.15	0.10	0.08
BRR with Fay 50%	1.02	1.01	1.01	1.03	0.20	0.15	0.10	0.08
Bootstrapping	0.99	0.99	1.00	1.02	0.22	0.17	0.12	0.10
JK2 Full	1.02	1.01	1.01	1.03	0.19	0.14	0.10	0.08
JK2 Half	1.02	1.01	1.01	1.03	0.19	0.14	0.10	0.08
JK2 Half with Fay 10%	1.02	1.01	1.01	1.03	0.19	0.14	0.10	0.08
JK2 Half with Fay 30%	1.02	1.01	1.01	1.03	0.19	0.14	0.10	0.08
JK2 Half with Fay 50%	1.02	1.01	1.01	1.03	0.19	0.14	0.10	0.08
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	1.01	1.04	0.98	1.06	0.19	0.16	0.10	0.08
BRR with Fay 10%	1.02	1.05	0.99	1.06	0.20	0.16	0.10	0.08
BRR with Fay 30%	1.06	1.06	0.99	1.07	0.21	0.16	0.10	0.08
BRR with Fay 50%	1.13	1.10	1.01	1.08	0.25	0.18	0.11	0.09
Bootstrapping	0.98	1.03	0.98	1.05	0.22	0.18	0.12	0.10
JK2 Full	1.01	1.05	0.99	1.06	0.19	0.15	0.10	0.08
JK2 Half	1.01	1.05	0.99	1.06	0.19	0.15	0.10	0.08
JK2 Half with Fay 10%	1.01	1.05	0.99	1.06	0.19	0.15	0.10	0.08
JK2 Half with Fay 30%	1.01	1.05	0.99	1.06	0.19	0.15	0.10	0.08
JK2 Half with Fay 50%	1.01	1.05	0.99	1.06	0.19	0.15	0.10	0.08
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.06	1.06	1.00	1.00	0.27	0.21	0.14	0.10
BRR with Fay 10%	1.06	1.06	1.00	1.00	0.27	0.21	0.14	0.10
BRR with Fay 30%	1.06	1.06	1.00	1.00	0.27	0.21	0.14	0.10
BRR with Fay 50%	1.06	1.06	1.00	1.00	0.27	0.21	0.14	0.10
BRR with Fay 50%: OrgStr	1.28	1.37	1.30	1.29	0.23	0.21	0.15	0.12
Bootstrapping	1.03	1.05	0.99	0.99	0.28	0.23	0.15	0.12
JK2 Full	1.06	1.06	1.00	1.00	0.25	0.20	0.13	0.10
JK2 Half	1.06	1.06	1.00	1.00	0.25	0.20	0.13	0.10
JK2 Half with Fay 10%	1.06	1.06	1.00	1.00	0.25	0.20	0.13	0.10
JK2 Half with Fay 30%	1.06	1.06	1.00	1.00	0.25	0.20	0.13	0.10
JK2 Half with Fay 50%	1.06	1.06	1.00	1.00	0.25	0.20	0.13	0.10
JK2 Half: OrgStr	0.81	0.80	0.78	0.78	0.17	0.13	0.09	0.07
JK2 Full: OrgStr	0.56	0.58	0.57	0.57	0.16	0.12	0.09	0.06

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 500 in all scenarios, with minor differences likely attributable to random sampling variation.

Table 6: Variance Estimator Comparison - Means Statistics (Sub-Group [Gender = Girl]) Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.01	1.00	1.01	1.05	0.21	0.16	0.11	0.09
BRR with Fay 10%	1.01	1.00	1.01	1.05	0.21	0.16	0.11	0.09
BRR with Fay 30%	1.01	1.00	1.01	1.05	0.21	0.15	0.11	0.09
BRR with Fay 50%	1.01	1.00	1.01	1.05	0.20	0.15	0.11	0.09
Bootstrapping	0.99	0.98	1.00	1.04	0.23	0.18	0.13	0.11
JK2 Full	1.01	1.00	1.01	1.05	0.20	0.15	0.11	0.09
JK2 Half	1.01	1.00	1.01	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 10%	1.01	1.00	1.01	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 30%	1.01	1.00	1.01	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 50%	1.01	1.00	1.01	1.05	0.20	0.15	0.11	0.09
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	1.00	1.00	0.99	1.04	0.21	0.16	0.11	0.09
BRR with Fay 10%	1.01	1.00	0.99	1.04	0.21	0.16	0.11	0.09
BRR with Fay 30%	1.04	1.02	1.00	1.05	0.23	0.17	0.11	0.09
BRR with Fay 50%	1.12	1.06	1.01	1.06	0.28	0.19	0.12	0.10
Bootstrapping	0.97	1.00	0.98	1.03	0.23	0.18	0.13	0.11
JK2 Full	1.00	1.01	0.99	1.05	0.20	0.15	0.11	0.09
JK2 Half	1.00	1.01	0.99	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 10%	1.00	1.01	0.99	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 30%	1.00	1.01	0.99	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 50%	1.00	1.01	0.99	1.05	0.20	0.15	0.11	0.09
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.06	1.04	0.97	1.01	0.31	0.24	0.15	0.13
BRR with Fay 10%	1.06	1.04	0.97	1.01	0.31	0.24	0.15	0.13
BRR with Fay 30%	1.06	1.04	0.97	1.01	0.30	0.24	0.15	0.13
BRR with Fay 50%	1.06	1.04	0.97	1.00	0.30	0.24	0.15	0.13
BRR with Fay 50%: OrgStr	1.25	1.31	1.24	1.27	0.26	0.23	0.16	0.13
Bootstrapping	1.03	1.02	0.96	0.99	0.31	0.25	0.16	0.14
JK2 Full	1.06	1.04	0.97	1.00	0.29	0.23	0.15	0.12
JK2 Half	1.06	1.04	0.97	1.00	0.29	0.23	0.15	0.12
JK2 Half with Fay 10%	1.06	1.04	0.97	1.00	0.29	0.23	0.15	0.12
JK2 Half with Fay 30%	1.06	1.04	0.97	1.00	0.29	0.23	0.15	0.12
JK2 Half with Fay 50%	1.06	1.04	0.97	1.00	0.29	0.23	0.15	0.12
JK2 Half: OrgStr	0.75	0.72	0.70	0.72	0.17	0.12	0.09	0.07
JK2 Full: OrgStr	0.52	0.52	0.51	0.53	0.16	0.12	0.08	0.06

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 504 for Score 1 and 513 for Score 2 in all scenarios, with minor differences likely attributable to random sampling variation.

Table 7: Variance Estimator Comparison - Difference in Means Statistics (Girls - Boys) Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	0.99	1.00	0.98	0.98	0.23	0.18	0.13	0.10
BRR with Fay 10%	0.99	1.00	0.98	0.98	0.23	0.18	0.13	0.10
BRR with Fay 30%	0.98	1.00	0.98	0.97	0.23	0.18	0.13	0.10
BRR with Fay 50%	0.98	1.00	0.98	0.97	0.23	0.18	0.13	0.10
Bootstrapping	0.96	0.99	0.97	0.97	0.24	0.20	0.14	0.11
JK2 Full	0.98	0.99	0.98	0.97	0.22	0.18	0.12	0.10
JK2 Half	0.98	0.99	0.98	0.97	0.22	0.18	0.12	0.10
JK2 Half with Fay 10%	0.98	0.99	0.98	0.97	0.22	0.18	0.12	0.10
JK2 Half with Fay 30%	0.98	0.99	0.98	0.97	0.22	0.18	0.12	0.10
JK2 Half with Fay 50%	0.98	0.99	0.98	0.97	0.22	0.18	0.12	0.10
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	0.93	0.95	0.98	0.97	0.22	0.18	0.12	0.10
BRR with Fay 10%	0.94	0.95	0.98	0.97	0.22	0.18	0.12	0.10
BRR with Fay 30%	0.96	0.96	0.99	0.98	0.23	0.18	0.13	0.10
BRR with Fay 50%	1.00	0.99	1.00	0.99	0.26	0.20	0.13	0.11
Bootstrapping	0.93	0.95	0.97	0.97	0.24	0.20	0.14	0.12
JK2 Full	0.97	0.97	0.99	0.98	0.22	0.18	0.12	0.10
JK2 Half	0.97	0.97	0.99	0.98	0.22	0.18	0.12	0.10
JK2 Half with Fay 10%	0.97	0.97	0.99	0.98	0.22	0.18	0.12	0.10
JK2 Half with Fay 30%	0.97	0.97	0.99	0.98	0.22	0.18	0.12	0.10
JK2 Half with Fay 50%	0.97	0.97	0.99	0.98	0.22	0.18	0.12	0.10
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	0.97	0.99	0.96	1.00	0.29	0.25	0.17	0.15
BRR with Fay 10%	0.97	0.99	0.96	0.99	0.29	0.25	0.17	0.15
BRR with Fay 30%	0.96	0.99	0.96	0.99	0.28	0.25	0.17	0.15
BRR with Fay 50%	0.96	0.99	0.96	0.99	0.28	0.25	0.17	0.15
BRR with Fay 50%: OrgStr	0.95	0.98	0.97	1.00	0.25	0.22	0.15	0.13
Bootstrapping	0.94	0.98	0.95	0.99	0.29	0.26	0.18	0.15
JK2 Full	0.96	0.99	0.96	0.99	0.27	0.24	0.16	0.14
JK2 Half	0.96	0.99	0.96	0.99	0.27	0.24	0.16	0.14
JK2 Half with Fay 10%	0.96	0.99	0.96	0.99	0.27	0.24	0.16	0.14
JK2 Half with Fay 30%	0.96	0.99	0.96	0.99	0.27	0.24	0.16	0.14
JK2 Half with Fay 50%	0.96	0.99	0.96	0.99	0.27	0.24	0.16	0.14
JK2 Half: OrgStr	0.73	0.71	0.70	0.73	0.19	0.15	0.10	0.09
JK2 Full: OrgStr	0.42	0.44	0.44	0.45	0.14	0.11	0.08	0.06

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 8 for Score 1 and 25 for Score 2 in all scenarios, with minor differences likely attributable to random sampling variation.

Table 8: Variance Estimator Comparison - 25th Percentile (Overall)

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.03	1.04	1.00	0.98	0.30	0.25	0.18	0.14
BRR with Fay 10%	1.03	1.04	1.00	0.98	0.31	0.26	0.18	0.15
BRR with Fay 30%	1.03	1.05	1.00	0.98	0.34	0.27	0.19	0.15
BRR with Fay 50%	1.04	1.05	1.01	0.98	0.38	0.30	0.21	0.17
Bootstrapping	1.02	1.04	0.99	0.97	0.32	0.26	0.18	0.15
JK2 Full	1.07	1.08	1.02	1.01	0.42	0.36	0.27	0.24
JK2 Half	1.06	1.08	1.02	1.01	0.43	0.37	0.27	0.24
JK2 Half with Fay 10%	1.06	1.08	1.02	1.01	0.44	0.38	0.28	0.25
JK2 Half with Fay 30%	1.07	1.09	1.03	1.02	0.47	0.41	0.30	0.28
JK2 Half with Fay 50%	1.09	1.10	1.04	1.03	0.52	0.46	0.34	0.32
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	1.02	1.04	0.99	1.02	0.30	0.26	0.17	0.15
BRR with Fay 10%	1.03	1.04	0.99	1.02	0.31	0.26	0.18	0.16
BRR with Fay 30%	1.06	1.06	1.00	1.02	0.33	0.28	0.19	0.17
BRR with Fay 50%	1.10	1.08	1.01	1.03	0.37	0.31	0.21	0.18
Bootstrapping	1.02	1.04	0.98	1.01	0.31	0.26	0.18	0.16
JK2 Full	1.05	1.07	1.02	1.05	0.40	0.36	0.28	0.26
JK2 Half	1.04	1.07	1.02	1.05	0.42	0.37	0.28	0.26
JK2 Half with Fay 10%	1.05	1.07	1.03	1.05	0.43	0.38	0.30	0.27
JK2 Half with Fay 30%	1.06	1.08	1.04	1.06	0.45	0.41	0.32	0.30
JK2 Half with Fay 50%	1.07	1.09	1.06	1.08	0.51	0.45	0.37	0.34
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.01	1.07	0.99	1.01	0.31	0.27	0.18	0.16
BRR with Fay 10%	1.02	1.07	0.99	1.01	0.32	0.27	0.19	0.16
BRR with Fay 30%	1.02	1.08	1.00	1.02	0.34	0.29	0.20	0.17
BRR with Fay 50%	1.03	1.08	1.00	1.02	0.37	0.31	0.22	0.18
BRR with Fay 50%: OrgStr	1.16	1.28	1.17	1.18	0.37	0.33	0.23	0.20
Bootstrapping	1.02	1.06	0.98	1.01	0.32	0.26	0.19	0.16
JK2 Full	1.05	1.10	1.03	1.04	0.40	0.36	0.28	0.26
JK2 Half	1.05	1.10	1.02	1.04	0.43	0.37	0.28	0.27
JK2 Half with Fay 10%	1.05	1.11	1.03	1.05	0.44	0.38	0.29	0.28
JK2 Half with Fay 30%	1.06	1.12	1.04	1.06	0.47	0.41	0.32	0.30
JK2 Half with Fay 50%	1.08	1.13	1.06	1.08	0.53	0.47	0.36	0.35
JK2 Half: OrgStr	0.96	1.02	0.95	0.97	0.40	0.35	0.27	0.25
JK2 Full: OrgStr	0.97	1.03	0.95	0.97	0.39	0.35	0.27	0.25

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 446 in all scenarios, with minor differences likely attributable to random sampling variation.

Table 9: Variance Estimator Comparison - Median (Overall)

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.01	1.01	1.04	1.02	0.29	0.22	0.17	0.14
BRR with Fay 10%	1.01	1.01	1.04	1.02	0.30	0.23	0.17	0.14
BRR with Fay 30%	1.02	1.01	1.04	1.02	0.32	0.24	0.18	0.15
BRR with Fay 50%	1.02	1.02	1.05	1.02	0.35	0.26	0.20	0.16
Bootstrapping	1.00	1.00	1.03	1.01	0.28	0.23	0.18	0.14
JK2 Full	1.04	1.04	1.07	1.04	0.39	0.31	0.27	0.24
JK2 Half	1.03	1.04	1.07	1.04	0.40	0.32	0.27	0.24
JK2 Half with Fay 10%	1.04	1.04	1.07	1.04	0.42	0.33	0.28	0.25
JK2 Half with Fay 30%	1.05	1.05	1.08	1.05	0.45	0.35	0.31	0.27
JK2 Half with Fay 50%	1.06	1.06	1.09	1.07	0.50	0.39	0.35	0.32
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	1.03	1.03	1.01	1.06	0.28	0.22	0.17	0.15
BRR with Fay 10%	1.04	1.04	1.01	1.06	0.29	0.23	0.17	0.15
BRR with Fay 30%	1.07	1.06	1.02	1.07	0.31	0.25	0.19	0.16
BRR with Fay 50%	1.14	1.10	1.04	1.08	0.37	0.28	0.21	0.17
Bootstrapping	1.00	1.02	1.01	1.05	0.28	0.23	0.18	0.16
JK2 Full	1.04	1.07	1.05	1.09	0.37	0.31	0.27	0.26
JK2 Half	1.04	1.07	1.05	1.09	0.37	0.33	0.27	0.26
JK2 Half with Fay 10%	1.04	1.08	1.05	1.10	0.38	0.33	0.28	0.27
JK2 Half with Fay 30%	1.05	1.08	1.06	1.11	0.41	0.36	0.31	0.30
JK2 Half with Fay 50%	1.06	1.10	1.08	1.13	0.44	0.41	0.36	0.35
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.04	1.04	1.02	1.04	0.31	0.26	0.19	0.16
BRR with Fay 10%	1.04	1.04	1.02	1.04	0.32	0.26	0.19	0.16
BRR with Fay 30%	1.04	1.04	1.02	1.04	0.34	0.28	0.20	0.17
BRR with Fay 50%	1.04	1.05	1.02	1.05	0.36	0.30	0.22	0.18
BRR with Fay 50%: OrgStr	1.22	1.31	1.27	1.29	0.37	0.32	0.23	0.20
Bootstrapping	1.01	1.02	1.01	1.03	0.31	0.26	0.19	0.16
JK2 Full	1.05	1.06	1.04	1.07	0.38	0.35	0.27	0.26
JK2 Half	1.05	1.06	1.04	1.07	0.40	0.36	0.27	0.26
JK2 Half with Fay 10%	1.05	1.06	1.04	1.08	0.41	0.36	0.28	0.27
JK2 Half with Fay 30%	1.05	1.07	1.05	1.09	0.43	0.39	0.30	0.30
JK2 Half with Fay 50%	1.06	1.08	1.07	1.10	0.48	0.44	0.34	0.34
JK2 Half: OrgStr	0.87	0.87	0.87	0.90	0.34	0.31	0.23	0.23
JK2 Full: OrgStr	0.87	0.88	0.87	0.90	0.33	0.30	0.23	0.23

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 498 to 500 in all scenarios, with differences likely attributable to random sampling variation.

Table 10: Variance Estimator Comparison - 75th Percentile (Overall)

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.02	1.02	1.06	1.07	0.30	0.24	0.18	0.16
BRR with Fay 10%	1.02	1.02	1.06	1.07	0.31	0.24	0.19	0.16
BRR with Fay 30%	1.02	1.02	1.06	1.07	0.33	0.26	0.20	0.17
BRR with Fay 50%	1.02	1.03	1.06	1.07	0.37	0.28	0.22	0.19
Bootstrapping	1.01	1.01	1.05	1.06	0.32	0.25	0.19	0.17
JK2 Full	1.04	1.06	1.08	1.10	0.40	0.34	0.29	0.27
JK2 Half	1.04	1.06	1.08	1.10	0.43	0.35	0.29	0.28
JK2 Half with Fay 10%	1.04	1.06	1.09	1.10	0.44	0.36	0.30	0.29
JK2 Half with Fay 30%	1.05	1.07	1.10	1.11	0.46	0.39	0.33	0.32
JK2 Half with Fay 50%	1.06	1.08	1.11	1.13	0.52	0.44	0.37	0.36
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	0.98	1.01	1.00	1.08	0.29	0.24	0.18	0.16
BRR with Fay 10%	0.99	1.02	1.00	1.08	0.30	0.25	0.18	0.17
BRR with Fay 30%	1.04	1.04	1.01	1.09	0.34	0.27	0.20	0.18
BRR with Fay 50%	1.14	1.09	1.03	1.10	0.43	0.31	0.22	0.20
Bootstrapping	0.98	1.02	1.00	1.08	0.30	0.25	0.18	0.17
JK2 Full	1.02	1.05	1.03	1.11	0.38	0.34	0.28	0.28
JK2 Half	1.02	1.05	1.03	1.11	0.39	0.35	0.29	0.28
JK2 Half with Fay 10%	1.02	1.05	1.03	1.11	0.40	0.36	0.30	0.29
JK2 Half with Fay 30%	1.04	1.06	1.04	1.12	0.44	0.38	0.33	0.32
JK2 Half with Fay 50%	1.05	1.07	1.06	1.14	0.49	0.43	0.37	0.37
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.12	1.05	1.01	1.02	0.42	0.34	0.22	0.18
BRR with Fay 10%	1.12	1.05	1.01	1.02	0.43	0.34	0.23	0.19
BRR with Fay 30%	1.13	1.06	1.02	1.02	0.45	0.36	0.24	0.20
BRR with Fay 50%	1.13	1.06	1.02	1.02	0.48	0.39	0.26	0.21
BRR with Fay 50%: OrgStr	1.32	1.34	1.29	1.29	0.47	0.40	0.28	0.24
Bootstrapping	1.11	1.04	1.01	1.01	0.39	0.34	0.23	0.19
JK2 Full	1.15	1.09	1.04	1.04	0.48	0.42	0.31	0.28
JK2 Half	1.13	1.08	1.04	1.04	0.51	0.43	0.31	0.28
JK2 Half with Fay 10%	1.14	1.08	1.04	1.05	0.52	0.44	0.32	0.29
JK2 Half with Fay 30%	1.15	1.09	1.05	1.05	0.56	0.46	0.34	0.31
JK2 Half with Fay 50%	1.16	1.10	1.07	1.07	0.60	0.50	0.38	0.35
JK2 Half: OrgStr	0.73	0.68	0.69	0.70	0.33	0.28	0.23	0.21
JK2 Full: OrgStr	0.73	0.68	0.69	0.70	0.32	0.26	0.22	0.21

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 554 in all scenarios, with minor differences likely attributable to random sampling variation.

Table 11: Variance Estimator Comparison - Low International Benchmark (Overall)

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
BRR with Fay 10%	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
BRR with Fay 30%	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
BRR with Fay 50%	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
Bootstrapping	0.99	0.99	0.95	0.99	0.30	0.23	0.16	0.13
JK2 Full	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
JK2 Half	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
JK2 Half with Fay 10%	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
JK2 Half with Fay 30%	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
JK2 Half with Fay 50%	1.01	1.01	0.96	1.00	0.28	0.21	0.15	0.12
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	0.99	1.04	0.99	0.99	0.28	0.23	0.15	0.12
BRR with Fay 10%	1.00	1.04	0.99	0.99	0.28	0.23	0.15	0.12
BRR with Fay 30%	1.01	1.05	0.99	0.99	0.28	0.23	0.15	0.12
BRR with Fay 50%	1.03	1.06	1.00	1.00	0.28	0.23	0.15	0.12
Bootstrapping	0.96	1.03	0.97	0.98	0.29	0.24	0.16	0.13
JK2 Full	0.99	1.04	0.99	0.99	0.27	0.22	0.15	0.12
JK2 Half	0.99	1.04	0.99	0.99	0.27	0.22	0.15	0.12
JK2 Half with Fay 10%	0.99	1.04	0.99	0.99	0.27	0.22	0.15	0.12
JK2 Half with Fay 30%	0.99	1.04	0.99	0.99	0.27	0.22	0.15	0.12
JK2 Half with Fay 50%	0.99	1.04	0.99	0.99	0.27	0.22	0.15	0.12
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	0.99	1.06	0.96	0.99	0.28	0.23	0.16	0.13
BRR with Fay 10%	0.99	1.06	0.96	0.99	0.28	0.23	0.16	0.13
BRR with Fay 30%	0.99	1.06	0.96	0.99	0.28	0.23	0.16	0.13
BRR with Fay 50%	0.99	1.06	0.96	0.99	0.28	0.23	0.16	0.13
BRR with Fay 50%: OrgStr	1.08	1.18	1.07	1.09	0.29	0.24	0.16	0.13
Bootstrapping	0.97	1.05	0.95	0.99	0.30	0.24	0.16	0.14
JK2 Full	0.99	1.06	0.96	1.00	0.28	0.22	0.15	0.12
JK2 Half	0.99	1.06	0.96	1.00	0.28	0.22	0.15	0.12
JK2 Half with Fay 10%	0.99	1.06	0.96	1.00	0.28	0.22	0.15	0.12
JK2 Half with Fay 30%	0.99	1.06	0.96	1.00	0.28	0.22	0.15	0.12
JK2 Half with Fay 50%	0.99	1.06	0.96	1.00	0.28	0.22	0.15	0.12
JK2 Half: OrgStr	0.96	1.03	0.93	0.96	0.27	0.22	0.15	0.12
JK2 Full: OrgStr	0.96	1.03	0.93	0.96	0.27	0.22	0.15	0.12

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 89% in all scenarios, with minor differences likely attributable to random sampling variation.

Table 12: Variance Estimator Comparison - High International Benchmark (Overall)

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.00	1.01	1.04	1.05	0.21	0.16	0.11	0.09
BRR with Fay 10%	1.00	1.01	1.04	1.05	0.21	0.16	0.11	0.09
BRR with Fay 30%	1.00	1.01	1.04	1.05	0.21	0.16	0.11	0.09
BRR with Fay 50%	1.00	1.01	1.04	1.05	0.21	0.16	0.11	0.09
Bootstrapping	0.98	1.00	1.03	1.04	0.23	0.18	0.13	0.10
JK2 Full	1.00	1.01	1.04	1.05	0.21	0.15	0.11	0.09
JK2 Half	1.00	1.01	1.04	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 10%	1.00	1.01	1.04	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 30%	1.00	1.01	1.04	1.05	0.20	0.15	0.11	0.09
JK2 Half with Fay 50%	1.00	1.01	1.04	1.05	0.20	0.15	0.11	0.09
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	0.98	0.99	0.99	1.07	0.20	0.15	0.11	0.09
BRR with Fay 10%	0.99	1.00	1.00	1.07	0.20	0.15	0.11	0.09
BRR with Fay 30%	1.04	1.02	1.00	1.08	0.22	0.16	0.11	0.09
BRR with Fay 50%	1.13	1.07	1.02	1.09	0.28	0.18	0.12	0.09
Bootstrapping	0.98	1.00	0.99	1.06	0.23	0.18	0.13	0.11
JK2 Full	1.00	1.02	1.00	1.08	0.20	0.15	0.11	0.09
JK2 Half	1.00	1.02	1.00	1.08	0.20	0.15	0.11	0.09
JK2 Half with Fay 10%	1.00	1.02	1.00	1.08	0.20	0.15	0.11	0.09
JK2 Half with Fay 30%	1.00	1.02	1.00	1.08	0.20	0.15	0.11	0.09
JK2 Half with Fay 50%	1.00	1.02	1.00	1.08	0.20	0.15	0.11	0.09
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.08	1.04	1.00	1.02	0.30	0.24	0.16	0.12
BRR with Fay 10%	1.08	1.04	1.00	1.02	0.30	0.24	0.16	0.12
BRR with Fay 30%	1.08	1.04	1.00	1.02	0.30	0.24	0.16	0.12
BRR with Fay 50%	1.08	1.04	1.00	1.02	0.30	0.24	0.16	0.12
BRR with Fay 50%: OrgStr	1.27	1.31	1.27	1.28	0.22	0.19	0.14	0.11
Bootstrapping	1.04	1.02	0.99	1.00	0.30	0.24	0.17	0.13
JK2 Full	1.07	1.04	1.00	1.02	0.28	0.23	0.16	0.12
JK2 Half	1.07	1.04	1.00	1.02	0.28	0.23	0.16	0.12
JK2 Half with Fay 10%	1.07	1.04	1.00	1.02	0.28	0.23	0.16	0.12
JK2 Half with Fay 30%	1.07	1.04	1.00	1.02	0.28	0.23	0.16	0.12
JK2 Half with Fay 50%	1.07	1.04	1.00	1.02	0.28	0.23	0.16	0.12
JK2 Half: OrgStr	0.71	0.68	0.68	0.69	0.18	0.12	0.09	0.07
JK2 Full: OrgStr	0.71	0.68	0.68	0.69	0.18	0.12	0.09	0.07

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 26% in all scenarios, with minor differences likely attributable to random sampling variation.

Table 13: Variance Estimator Comparison - Advanced International Benchmark (Overall)

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	0.96	0.93	1.00	1.05	0.35	0.25	0.19	0.17
BRR with Fay 10%	0.96	0.93	1.00	1.05	0.35	0.25	0.19	0.17
BRR with Fay 30%	0.96	0.93	1.00	1.05	0.35	0.25	0.19	0.17
BRR with Fay 50%	0.96	0.93	1.00	1.05	0.35	0.25	0.19	0.17
Bootstrapping	0.94	0.91	0.99	1.05	0.35	0.26	0.20	0.18
JK2 Full	0.96	0.93	1.00	1.05	0.34	0.25	0.19	0.17
JK2 Half	0.96	0.93	1.00	1.05	0.34	0.25	0.19	0.17
JK2 Half with Fay 10%	0.96	0.93	1.00	1.05	0.34	0.25	0.19	0.17
JK2 Half with Fay 30%	0.96	0.93	1.00	1.05	0.34	0.25	0.19	0.17
JK2 Half with Fay 50%	0.96	0.93	1.00	1.05	0.34	0.25	0.19	0.17
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	0.87	0.96	0.97	1.02	0.31	0.27	0.19	0.16
BRR with Fay 10%	0.89	0.97	0.97	1.02	0.32	0.27	0.19	0.16
BRR with Fay 30%	0.93	0.99	0.98	1.02	0.35	0.27	0.19	0.16
BRR with Fay 50%	1.03	1.03	0.99	1.03	0.44	0.30	0.19	0.16
Bootstrapping	0.90	0.97	0.96	1.01	0.35	0.28	0.20	0.17
JK2 Full	0.93	0.99	0.98	1.02	0.34	0.27	0.19	0.16
JK2 Half	0.93	0.99	0.98	1.02	0.34	0.27	0.19	0.16
JK2 Half with Fay 10%	0.93	0.99	0.98	1.02	0.34	0.27	0.19	0.16
JK2 Half with Fay 30%	0.93	0.99	0.98	1.02	0.34	0.27	0.19	0.16
JK2 Half with Fay 50%	0.93	0.99	0.98	1.02	0.34	0.27	0.19	0.16
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.02	0.97	0.92	0.89	0.51	0.40	0.27	0.22
BRR with Fay 10%	1.02	0.97	0.92	0.89	0.51	0.40	0.27	0.22
BRR with Fay 30%	1.02	0.97	0.92	0.89	0.50	0.40	0.27	0.22
BRR with Fay 50%	1.02	0.97	0.92	0.89	0.50	0.40	0.27	0.22
BRR with Fay 50%: OrgStr	1.10	1.11	1.09	1.05	0.45	0.39	0.29	0.23
Bootstrapping	0.98	0.96	0.92	0.88	0.49	0.40	0.28	0.22
JK2 Full	1.02	0.97	0.93	0.89	0.50	0.40	0.27	0.22
JK2 Half	1.02	0.97	0.93	0.89	0.50	0.40	0.27	0.22
JK2 Half with Fay 10%	1.02	0.97	0.93	0.89	0.50	0.40	0.27	0.22
JK2 Half with Fay 30%	1.02	0.97	0.93	0.89	0.50	0.40	0.27	0.22
JK2 Half with Fay 50%	1.02	0.97	0.93	0.89	0.50	0.40	0.27	0.22
JK2 Half: OrgStr	0.41	0.37	0.38	0.37	0.20	0.14	0.11	0.08
JK2 Full: OrgStr	0.41	0.37	0.38	0.37	0.20	0.14	0.11	0.08

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 6% in all scenarios, with minor differences likely attributable to random sampling variation.

Table 14: Variance Estimator Comparison - Pearson Correlation between Score 1 and Score 2 (Overall)

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.03	1.00	1.00	1.01	0.26	0.21	0.14	0.12
BRR with Fay 10%	1.02	0.99	1.00	1.01	0.25	0.20	0.14	0.12
BRR with Fay 30%	1.00	0.98	1.00	1.00	0.24	0.19	0.14	0.12
BRR with Fay 50%	0.99	0.98	0.99	1.00	0.23	0.19	0.14	0.12
Bootstrapping	0.96	0.92	0.92	0.92	0.25	0.18	0.13	0.12
JK2 Full	0.98	0.97	0.99	1.00	0.23	0.18	0.14	0.12
JK2 Half	0.98	0.97	0.99	1.00	0.23	0.18	0.13	0.12
JK2 Half with Fay 10%	0.98	0.97	0.99	1.00	0.23	0.18	0.13	0.12
JK2 Half with Fay 30%	0.98	0.97	0.99	1.00	0.23	0.18	0.13	0.12
JK2 Half with Fay 50%	0.98	0.97	0.99	1.00	0.23	0.18	0.13	0.12
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	0.93	1.00	0.97	0.99	0.24	0.21	0.15	0.12
BRR with Fay 10%	0.93	0.99	0.97	0.99	0.23	0.21	0.15	0.12
BRR with Fay 30%	0.94	0.99	0.97	0.99	0.24	0.20	0.14	0.11
BRR with Fay 50%	0.98	1.01	0.97	0.99	0.27	0.20	0.14	0.11
Bootstrapping	0.91	0.92	0.90	0.90	0.24	0.19	0.14	0.11
JK2 Full	0.93	0.99	0.96	0.98	0.22	0.19	0.14	0.11
JK2 Half	0.93	0.99	0.96	0.98	0.23	0.20	0.14	0.11
JK2 Half with Fay 10%	0.93	0.99	0.96	0.98	0.23	0.20	0.14	0.11
JK2 Half with Fay 30%	0.92	0.98	0.96	0.98	0.22	0.19	0.14	0.11
JK2 Half with Fay 50%	0.92	0.98	0.96	0.98	0.22	0.19	0.14	0.11
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.07	1.03	0.91	0.94	0.36	0.29	0.18	0.17
BRR with Fay 10%	1.05	1.02	0.91	0.94	0.34	0.28	0.18	0.17
BRR with Fay 30%	1.03	1.01	0.90	0.93	0.32	0.27	0.17	0.17
BRR with Fay 50%	1.01	1.00	0.90	0.93	0.31	0.26	0.17	0.16
BRR with Fay 50%: OrgStr	1.01	1.03	0.96	1.00	0.25	0.21	0.17	0.15
Bootstrapping	0.98	0.94	0.83	0.85	0.32	0.24	0.16	0.15
JK2 Full	1.02	0.99	0.90	0.93	0.31	0.26	0.17	0.16
JK2 Half	1.01	0.99	0.90	0.92	0.33	0.26	0.17	0.16
JK2 Half with Fay 10%	1.01	0.99	0.90	0.92	0.32	0.26	0.17	0.16
JK2 Half with Fay 30%	1.00	0.99	0.89	0.92	0.31	0.25	0.17	0.16
JK2 Half with Fay 50%	1.00	0.98	0.89	0.92	0.30	0.25	0.17	0.16
JK2 Half: OrgStr	0.65	0.63	0.58	0.60	0.19	0.15	0.10	0.08
JK2 Full: OrgStr	0.65	0.63	0.58	0.60	0.19	0.15	0.10	0.08

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 0.34 in all scenarios, with minor differences likely attributable to random sampling variation.

Table 15: Variance Estimator Comparison - All Statistics (Overall and Sub-Group (Girls))

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.01	1.01	1.02	1.03	0.28	0.22	0.16	0.13
BRR with Fay 10%	1.01	1.01	1.02	1.03	0.28	0.22	0.16	0.13
BRR with Fay 30%	1.01	1.01	1.02	1.03	0.29	0.23	0.17	0.14
BRR with Fay 50%	1.02	1.01	1.02	1.03	0.31	0.24	0.17	0.15
Bootstrapping	0.99	0.99	1.01	1.02	0.29	0.23	0.17	0.14
JK2 Full	1.03	1.02	1.03	1.05	0.33	0.27	0.21	0.19
JK2 Half	1.02	1.02	1.03	1.05	0.34	0.27	0.22	0.19
JK2 Half with Fay 10%	1.02	1.02	1.04	1.05	0.34	0.28	0.22	0.20
JK2 Half with Fay 30%	1.03	1.03	1.04	1.06	0.36	0.29	0.24	0.22
JK2 Half with Fay 50%	1.04	1.04	1.05	1.07	0.39	0.32	0.26	0.25
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	0.98	1.00	0.99	1.03	0.27	0.22	0.16	0.13
BRR with Fay 10%	0.99	1.01	1.00	1.04	0.27	0.22	0.16	0.14
BRR with Fay 30%	1.02	1.03	1.00	1.04	0.30	0.24	0.17	0.14
BRR with Fay 50%	1.10	1.06	1.02	1.05	0.35	0.26	0.18	0.15
Bootstrapping	0.97	1.00	0.98	1.02	0.28	0.23	0.17	0.14
JK2 Full	1.01	1.04	1.02	1.06	0.32	0.27	0.21	0.20
JK2 Half	1.01	1.03	1.02	1.05	0.33	0.28	0.22	0.20
JK2 Half with Fay 10%	1.01	1.04	1.02	1.06	0.33	0.28	0.22	0.21
JK2 Half with Fay 30%	1.01	1.04	1.02	1.06	0.35	0.30	0.24	0.22
JK2 Half with Fay 50%	1.02	1.05	1.03	1.07	0.38	0.33	0.27	0.25
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.05	1.04	0.99	1.00	0.35	0.28	0.19	0.16
BRR with Fay 10%	1.05	1.04	0.99	1.00	0.35	0.29	0.19	0.16
BRR with Fay 30%	1.05	1.04	0.99	1.01	0.36	0.30	0.20	0.17
BRR with Fay 50%	1.05	1.04	0.99	1.01	0.38	0.31	0.21	0.17
BRR with Fay 50%: OrgStr	1.19	1.24	1.18	1.20	0.35	0.31	0.22	0.18
Bootstrapping	1.02	1.02	0.97	0.99	0.34	0.28	0.20	0.16
JK2 Full	1.06	1.05	1.00	1.02	0.38	0.32	0.24	0.21
JK2 Half	1.05	1.05	1.00	1.02	0.39	0.33	0.24	0.22
JK2 Half with Fay 10%	1.06	1.05	1.00	1.02	0.40	0.34	0.25	0.22
JK2 Half with Fay 30%	1.06	1.06	1.01	1.03	0.41	0.35	0.26	0.24
JK2 Half with Fay 50%	1.07	1.07	1.02	1.04	0.44	0.38	0.28	0.26
JK2 Half: OrgStr	0.78	0.77	0.75	0.77	0.28	0.23	0.18	0.17
JK2 Full: OrgStr	0.72	0.72	0.70	0.71	0.27	0.22	0.17	0.16

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. Statistic types include means (overall and sub-group level (girls)), difference in means between girls and boys, 25th, 50th and 75th percentiles (overall and sub-group level (girls)), low, high, and advanced benchmarks (overall), and Pearson correlations (overall) between Score 1 and Score 2 across all sampling scenarios.

11 Figures

Figure 1

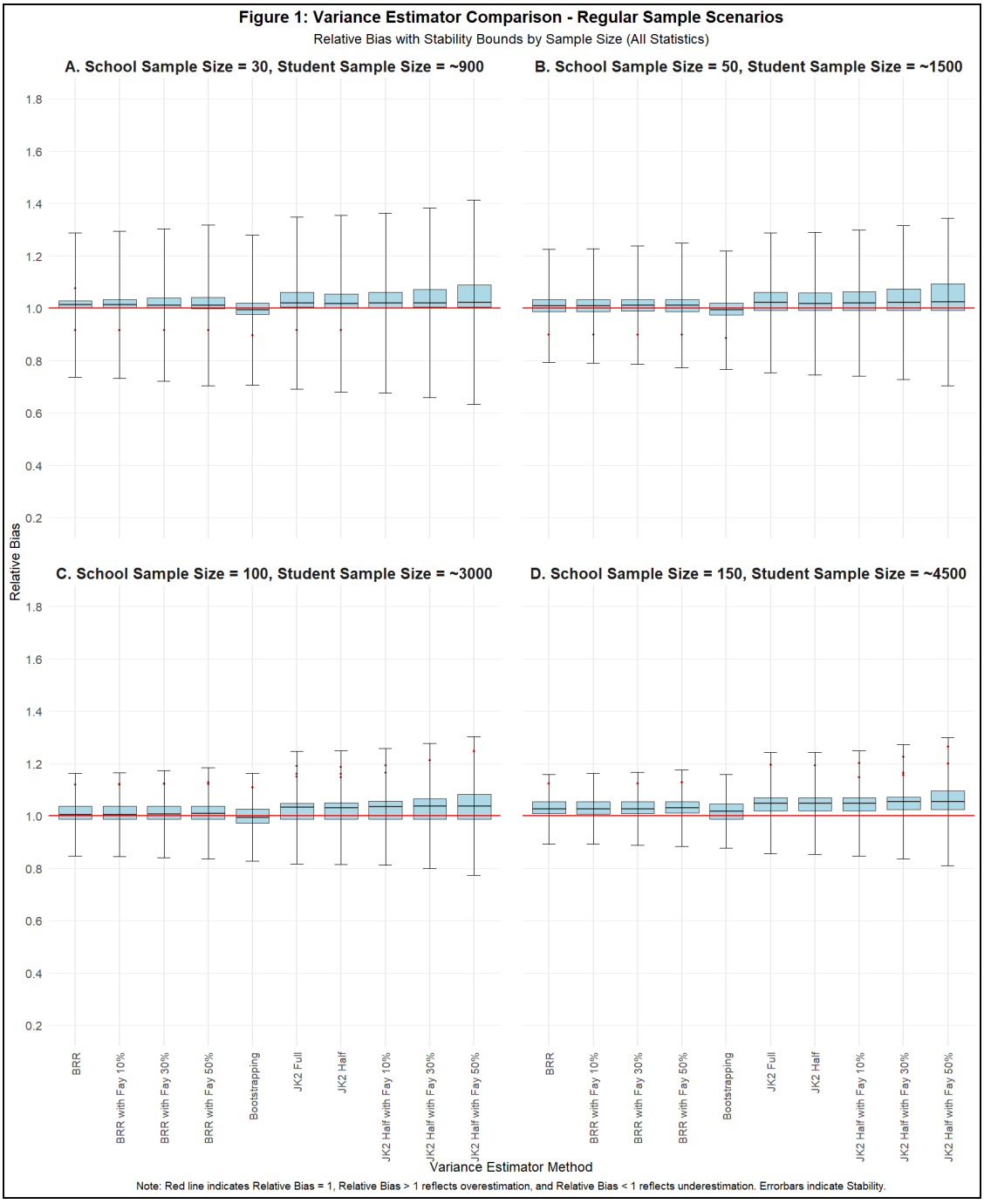

Figure 2

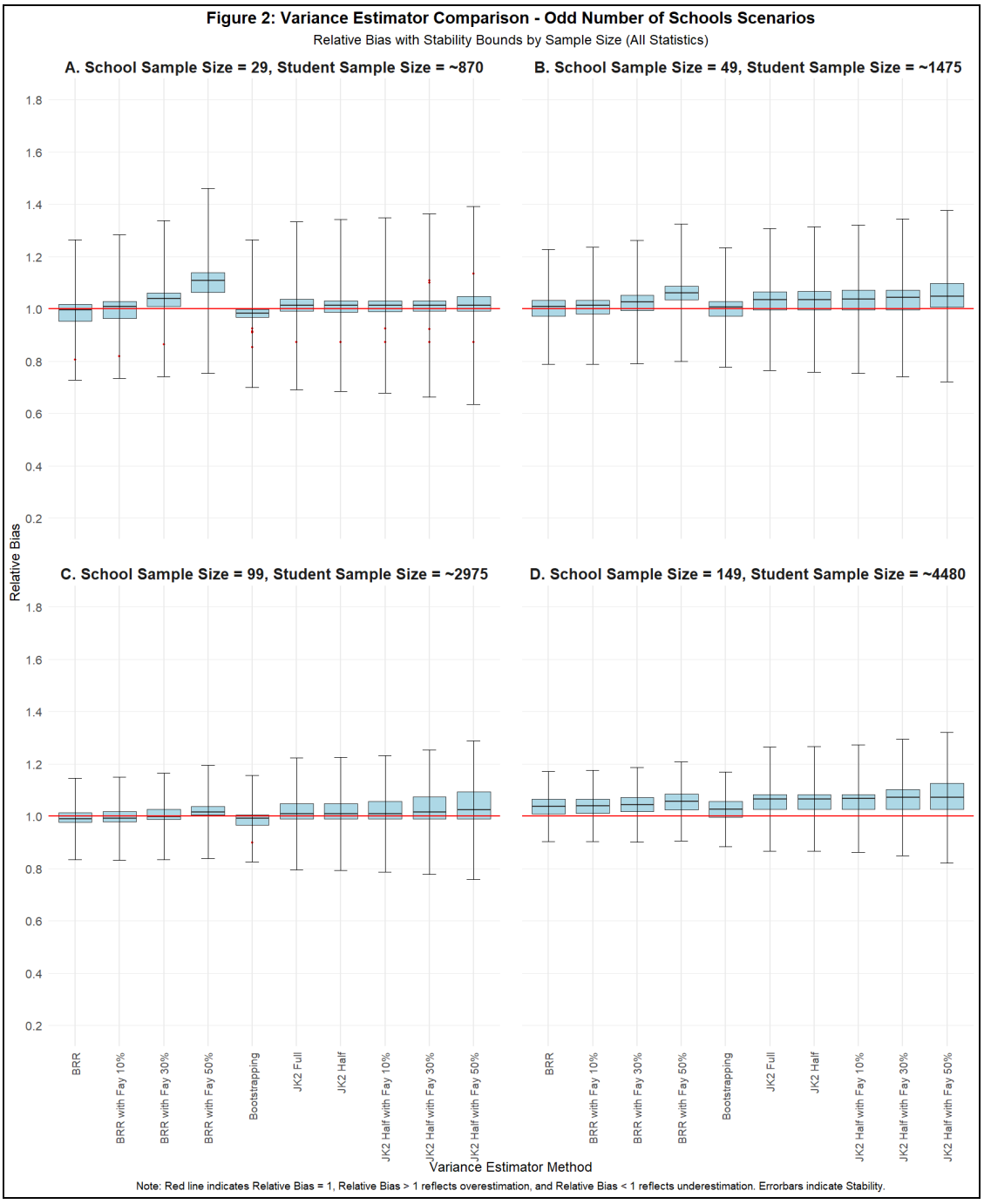
Figure 2: Variance Estimator Comparison - Odd Number of Schools Scenarios
Relative Bias with Stability Bounds by Sample Size (All Statistics)
A. School Sample Size = 29, Student Sample Size = ~870
B. School Sample Size = 49, Student Sample Size = ~1475
C. School Sample Size = 99, Student Sample Size = ~2975
D. School Sample Size = 149, Student Sample Size = ~4480
Relative Bias
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
BRR
BRR with Fay 10%
BRR with Fay 30%
BRR with Fay 50%
Bootstrapping
JK2 Full
JK2 Half
JK2 Half with Fay 10%
JK2 Half with Fay 30%
JK2 Half with Fay 50%
Variance Estimator Method
Note: Red line indicates Relative Bias = 1, Relative Bias > 1 reflects overestimation, and Relative Bias < 1 reflects underestimation. Errorbars indicate Stability.

Figure 3

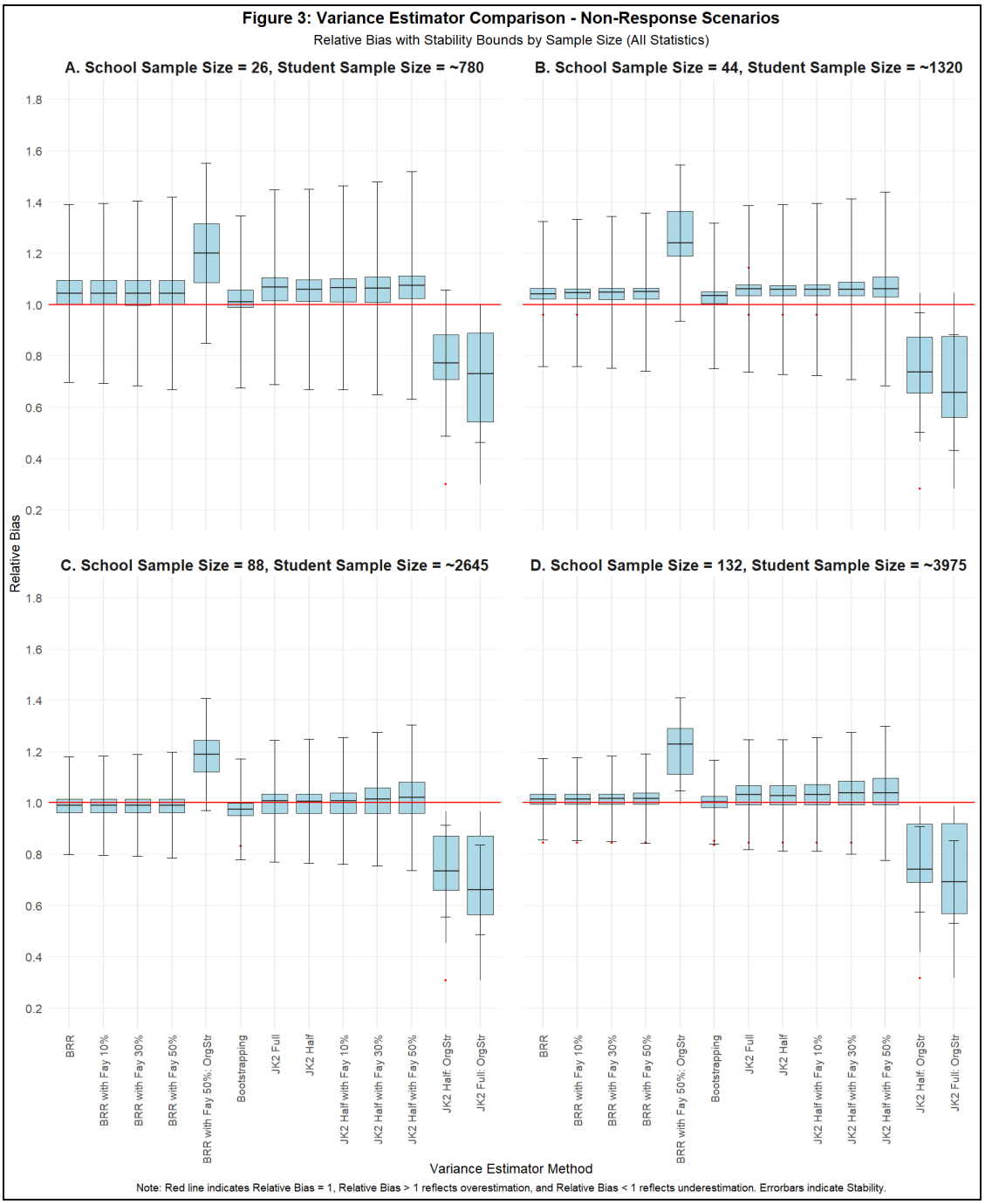
Figure 3: Variance Estimator Comparison - Non-Response Scenarios
Relative Bias with Stability Bounds by Sample Size (All Statistics)
A. School Sample Size = 26, Student Sample Size = ~780
B. School Sample Size = 44, Student Sample Size = ~1320
C. School Sample Size = 88, Student Sample Size = ~2645
D. School Sample Size = 132, Student Sample Size = ~3975
Relative Bias
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
BRR
BRR with Fay 10%
BRR with Fay 30%
BRR with Fay 50%
BRR with Fay 50%: OrgStr
Bootstrapping
JK2 Full
JK2 Half
JK2 Half with Fay 10%
JK2 Half with Fay 30%
JK2 Half with Fay 50%
JK2 Half: OrgStr
JK2 Full: OrgStr
Variance Estimator Method
Note: Red line indicates Relative Bias = 1, Relative Bias > 1 reflects overestimation, and Relative Bias < 1 reflects underestimation. Errorbars indicate Stability.

12 Appendix A: TIMSS 2023 Statistics

Table Appendix A 1: Mathematics Achievement Statistics - TIMSS 2023

Country	Mean	Standard Deviation (SD)	Intra-Class Coefficient (ICC)
Albania	509	81	0.31
Azerbaijan, Republic of	505	89	0.28
Australia	521	91	0.24
Bahrain	457	100	0.41
Armenia	512	69	0.21
Bosnia and Herzegovina	451	72	0.21
Brazil	377	83	0.36
Bulgaria	545	87	0.46
Canada	496	81	0.22
Chile	451	81	0.22
Chinese Taipei	604	67	0.11
Cyprus	518	81	0.13
Czech Republic	527	74	0.22
Denmark	525	71	0.14
Finland	530	80	0.15
France	483	77	0.16
Georgia	498	77	0.29
Germany	523	75	0.27
Hong Kong, SAR	593	83	0.34
Hungary	529	90	0.39
Iran, Islamic Republic of	420	97	0.37
Ireland	547	81	0.12
Italy	514	80	0.17
Japan	593	71	0.09
Kazakhstan	488	86	0.29
Jordan	417	100	0.40
Korea, Republic of	598	74	0.18
Kosovo	459	79	0.29
Kuwait	384	115	0.38
Latvia	539	76	0.12
Lithuania	563	76	0.26
Macao SAR	583	79	0.22
Montenegro	477	76	0.20
Morocco	384	99	0.43
Oman	423	98	0.46
Netherlands	539	65	0.14
New Zealand	499	91	0.19
Norway	532	74	0.10
Poland	547	76	0.09
Portugal	517	82	0.24
Qatar	465	97	0.36
Romania	551	81	0.30
Saudi Arabia	417	93	0.36
Serbia	528	78	0.29
Singapore	611	88	0.16
Slovak Republic	516	82	0.36
Slovenia	515	73	0.05
South Africa	353	109	0.50
Spain	504	75	0.18
Sweden	534	75	0.15

Table Appendix A 1: Mathematics Achievement Statistics - TIMSS 2023

Country	Mean	Standard Deviation (SD)	Intra-Class Coefficient (ICC)
United Arab Emirates	496	104	0.49
Turkey	552	100	0.28
North Macedonia	472	86	0.26
United States	517	98	0.25
Uzbekistan	445	82	0.14
England	551	91	0.14
Belgium (Flemish)	524	70	0.18
Belgium (French)	491	75	0.18
United Arab Emirates (Dubai)	550	94	0.41
United Arab Emirates (Abu Dhabi)	468	106	0.46
United Arab Emirates (Sharjah)	498	98	0.43
Canada (Ontario)	503	82	0.19
Canada (Quebec)	517	74	0.13
Average	504	84	0.26

Note: Statistics are based on TIMSS 2023 mathematics achievement scores using the first plausible value. ICC explains the variation in mathematics achievement at the school level.

13 Appendix B: Percentiles by Sub-Group

Appendix B 1: Variance Estimator Comparison - 25th Percentile (Sub-Group [Gender = Girl])

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.03	1.04	1.00	1.01	0.31	0.26	0.18	0.16
BRR with Fay 10%	1.04	1.05	1.00	1.02	0.33	0.27	0.18	0.17
BRR with Fay 30%	1.04	1.05	1.00	1.02	0.35	0.29	0.20	0.18
BRR with Fay 50%	1.06	1.06	1.01	1.03	0.39	0.32	0.22	0.20
Bootstrapping	1.01	1.03	0.99	1.00	0.31	0.26	0.18	0.16
JK2 Full	1.10	1.09	1.06	1.08	0.46	0.41	0.33	0.32
JK2 Half	1.09	1.09	1.05	1.08	0.49	0.42	0.33	0.33
JK2 Half with Fay 10%	1.10	1.09	1.06	1.09	0.51	0.44	0.35	0.35
JK2 Half with Fay 30%	1.12	1.11	1.08	1.10	0.56	0.48	0.39	0.39
JK2 Half with Fay 50%	1.14	1.14	1.11	1.13	0.63	0.56	0.46	0.46
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	1.01	1.03	1.00	1.02	0.31	0.26	0.18	0.17
BRR with Fay 10%	1.01	1.03	1.00	1.02	0.32	0.27	0.19	0.17
BRR with Fay 30%	1.04	1.05	1.01	1.03	0.35	0.30	0.20	0.18
BRR with Fay 50%	1.10	1.09	1.03	1.04	0.39	0.33	0.23	0.20
Bootstrapping	1.00	1.02	0.99	1.01	0.31	0.26	0.19	0.17
JK2 Full	1.06	1.09	1.06	1.08	0.44	0.41	0.33	0.34
JK2 Half	1.04	1.09	1.06	1.07	0.45	0.42	0.34	0.34
JK2 Half with Fay 10%	1.05	1.10	1.06	1.08	0.47	0.44	0.35	0.36
JK2 Half with Fay 30%	1.06	1.11	1.07	1.10	0.51	0.49	0.39	0.40
JK2 Half with Fay 50%	1.09	1.15	1.10	1.12	0.58	0.57	0.46	0.48
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	0.99	1.05	1.00	1.04	0.31	0.27	0.20	0.18
BRR with Fay 10%	0.99	1.05	1.00	1.05	0.32	0.28	0.20	0.18
BRR with Fay 30%	1.00	1.06	1.00	1.05	0.35	0.30	0.22	0.19
BRR with Fay 50%	1.01	1.06	1.01	1.05	0.38	0.33	0.24	0.21
BRR with Fay 50%: OrgStr	1.14	1.25	1.17	1.21	0.39	0.36	0.25	0.22
Bootstrapping	0.99	1.04	0.98	1.03	0.32	0.27	0.19	0.17
JK2 Full	1.04	1.09	1.04	1.10	0.43	0.41	0.34	0.34
JK2 Half	1.03	1.08	1.04	1.10	0.45	0.42	0.35	0.35
JK2 Half with Fay 10%	1.04	1.09	1.05	1.10	0.46	0.44	0.36	0.37
JK2 Half with Fay 30%	1.04	1.10	1.06	1.12	0.51	0.48	0.40	0.41
JK2 Half with Fay 50%	1.06	1.13	1.09	1.15	0.57	0.56	0.47	0.47
JK2 Half: OrgStr	0.92	0.96	0.93	0.98	0.42	0.39	0.32	0.33
JK2 Full: OrgStr	0.93	0.97	0.93	0.98	0.40	0.38	0.31	0.32

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 450 for Score 1 and 460 for Score 2 in all scenarios, with minor differences likely attributable to random sampling variation.

Appendix B 2: Variance Estimator Comparison - Median (Sub-Group [Gender = Girl])

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.05	1.03	1.05	1.04	0.32	0.25	0.20	0.16
BRR with Fay 10%	1.05	1.03	1.05	1.04	0.33	0.26	0.20	0.16
BRR with Fay 30%	1.04	1.03	1.05	1.04	0.35	0.27	0.22	0.17
BRR with Fay 50%	1.05	1.03	1.06	1.04	0.38	0.29	0.23	0.18
Bootstrapping	1.02	1.02	1.04	1.03	0.32	0.26	0.20	0.16
JK2 Full	1.07	1.05	1.09	1.08	0.44	0.37	0.34	0.31
JK2 Half	1.07	1.05	1.09	1.08	0.46	0.38	0.35	0.31
JK2 Half with Fay 10%	1.07	1.05	1.09	1.09	0.47	0.39	0.36	0.33
JK2 Half with Fay 30%	1.08	1.07	1.11	1.11	0.51	0.42	0.40	0.37
JK2 Half with Fay 50%	1.10	1.09	1.13	1.14	0.58	0.49	0.46	0.44
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	1.02	1.00	1.03	1.05	0.31	0.25	0.19	0.16
BRR with Fay 10%	1.03	1.01	1.03	1.05	0.32	0.26	0.20	0.17
BRR with Fay 30%	1.06	1.02	1.04	1.05	0.35	0.28	0.21	0.18
BRR with Fay 50%	1.14	1.07	1.06	1.07	0.42	0.32	0.24	0.20
Bootstrapping	1.00	0.99	1.02	1.04	0.31	0.25	0.20	0.17
JK2 Full	1.04	1.05	1.08	1.09	0.43	0.38	0.34	0.31
JK2 Half	1.03	1.05	1.08	1.09	0.44	0.40	0.34	0.31
JK2 Half with Fay 10%	1.03	1.06	1.08	1.09	0.46	0.42	0.36	0.32
JK2 Half with Fay 30%	1.04	1.07	1.10	1.11	0.49	0.46	0.40	0.36
JK2 Half with Fay 50%	1.06	1.10	1.12	1.13	0.56	0.52	0.46	0.43
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.09	1.07	1.04	1.06	0.37	0.30	0.22	0.19
BRR with Fay 10%	1.09	1.07	1.04	1.06	0.38	0.31	0.22	0.19
BRR with Fay 30%	1.10	1.07	1.04	1.06	0.41	0.33	0.23	0.20
BRR with Fay 50%	1.10	1.07	1.04	1.06	0.45	0.36	0.25	0.21
BRR with Fay 50%: OrgStr	1.27	1.32	1.27	1.29	0.44	0.38	0.27	0.23
Bootstrapping	1.07	1.05	1.03	1.04	0.39	0.30	0.23	0.19
JK2 Full	1.12	1.11	1.07	1.10	0.47	0.40	0.33	0.32
JK2 Full: OrgStr	0.86	0.85	0.82	0.86	0.38	0.33	0.28	0.27
JK2 Half	1.11	1.10	1.07	1.10	0.49	0.41	0.34	0.32
JK2 Half with Fay 10%	1.11	1.10	1.07	1.11	0.50	0.43	0.35	0.33
JK2 Half with Fay 30%	1.12	1.12	1.08	1.12	0.54	0.47	0.38	0.37
JK2 Half with Fay 50%	1.14	1.14	1.11	1.15	0.62	0.53	0.43	0.44
JK2 Half: OrgStr	0.86	0.84	0.82	0.86	0.40	0.34	0.28	0.28

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 504 for Score 1 and 515 for Score 2 in all scenarios, with minor differences likely attributable to random sampling variation.

Appendix B 3: Variance Estimator Comparison - 75th Percentile (Sub-Group [Gender = Girl])

Relative Bias and Stability Across Sampling Scenarios

Variance Estimator Method	Relative Bias				Stability			
	Very Small Samples	Small Samples	Medium Samples	Large Samples	Very Small Samples	Small Samples	Medium Samples	Large Samples
1. Regular Sample Scenario (Schools: 30 for Very Small, 50 for Small, 100 for Medium, and 150 for Large)								
BRR	1.02	1.00	1.07	1.08	0.32	0.26	0.21	0.18
BRR with Fay 10%	1.02	1.00	1.07	1.08	0.33	0.27	0.21	0.19
BRR with Fay 30%	1.02	1.00	1.07	1.08	0.36	0.29	0.23	0.20
BRR with Fay 50%	1.02	1.01	1.07	1.09	0.39	0.32	0.25	0.22
Bootstrapping	1.01	0.99	1.06	1.07	0.34	0.27	0.21	0.19
JK2 Full	1.05	1.04	1.11	1.13	0.44	0.39	0.35	0.32
JK2 Half	1.03	1.04	1.11	1.13	0.46	0.40	0.35	0.33
JK2 Half with Fay 10%	1.04	1.04	1.11	1.14	0.48	0.42	0.37	0.34
JK2 Half with Fay 30%	1.05	1.05	1.13	1.16	0.52	0.45	0.41	0.38
JK2 Half with Fay 50%	1.08	1.08	1.15	1.18	0.59	0.52	0.47	0.45
2. Odd Number of Schools Scenario (Schools: 29 for Very Small, 49 for Small, 99 for Medium, and 149 for Large)								
BRR	0.98	0.96	1.00	1.06	0.32	0.26	0.20	0.18
BRR with Fay 10%	0.99	0.97	1.00	1.06	0.34	0.27	0.20	0.19
BRR with Fay 30%	1.04	0.99	1.01	1.07	0.39	0.29	0.22	0.20
BRR with Fay 50%	1.14	1.04	1.04	1.09	0.50	0.34	0.24	0.23
Bootstrapping	1.00	0.97	0.99	1.05	0.35	0.27	0.21	0.18
JK2 Full	1.05	1.02	1.04	1.11	0.45	0.38	0.33	0.34
JK2 Half	1.04	1.01	1.04	1.11	0.49	0.39	0.33	0.34
JK2 Half with Fay 10%	1.05	1.02	1.05	1.12	0.50	0.41	0.35	0.35
JK2 Half with Fay 30%	1.06	1.03	1.06	1.13	0.55	0.45	0.38	0.39
JK2 Half with Fay 50%	1.07	1.05	1.09	1.16	0.61	0.51	0.45	0.46
3. Non-Response Scenario (Schools: 26 for Very Small, 44 for Small, 88 for Medium, and 132 for Large)								
BRR	1.13	1.03	0.99	1.02	0.47	0.38	0.25	0.21
BRR with Fay 10%	1.13	1.04	0.99	1.02	0.48	0.38	0.25	0.22
BRR with Fay 30%	1.14	1.04	0.99	1.02	0.52	0.41	0.27	0.23
BRR with Fay 50%	1.15	1.05	0.99	1.02	0.56	0.44	0.29	0.24
BRR with Fay 50%: OrgStr	1.29	1.25	1.20	1.24	0.53	0.43	0.32	0.26
Bootstrapping	1.11	1.04	0.98	1.01	0.45	0.38	0.25	0.21
JK2 Full	1.17	1.08	1.03	1.07	0.57	0.47	0.35	0.32
JK2 Half	1.14	1.06	1.02	1.07	0.60	0.49	0.37	0.32
JK2 Half with Fay 10%	1.15	1.06	1.03	1.07	0.61	0.50	0.38	0.33
JK2 Half with Fay 30%	1.16	1.08	1.04	1.09	0.66	0.53	0.41	0.37
JK2 Half with Fay 50%	1.18	1.10	1.07	1.11	0.74	0.59	0.46	0.42
JK2 Half: OrgStr	0.68	0.61	0.63	0.65	0.37	0.29	0.25	0.24
JK2 Full: OrgStr	0.68	0.61	0.63	0.66	0.34	0.28	0.25	0.23

Note: **Relative Bias** = 1 indicates no bias in estimation of sampling errors, **Relative Bias** > 1 reflects overestimation, and **Relative Bias** < 1 reflects underestimation. **Stability** values closer to 0 represent higher precision of the sampling error estimate. Highlighted cells indicate the best-performing method(s) per sample size scenario. The average of the estimates across 1,000 samples is approximately 558 for Score 1 and 565 for Score 2 in all scenarios, with minor differences likely attributable to random sampling variation.

14 Declarations

Acknowledgements

We thank the International Association for the Evaluation of Educational Achievement (IEA) for funding this project through the IEA's third round of Research and Development Funds (R&D). We also gratefully acknowledge the anonymous reviewer whose invaluable feedback helped us substantially improve the quality of this report.

Funding

This work was supported by the International Association for the Evaluation of Educational Achievement (IEA) under the third round of the Research and Development Fund (Grant title: An Examination of the Performance of Sampling Variance Estimators in International Large-Scale Assessments, 2024; EUR 50,000).

Author Information

Affiliations: International Association for the Evaluation of Educational Achievement (IEA), Überseering 27, 22297, Hamburg, Germany

Authors: Umut Atasever, Sabine Meinck & Diego Cortes

Corresponding author

Correspondence to Umut Atasever.