Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Umut Atasever

umut.atasever@iea-hamburg.de

IEA Hamburg

Sabine Meinck

IEA Hamburg

Diego Cortes

IEA Hamburg

Method Article

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

Posted Date: August 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10472250/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

2026.07.24

Abstract

Background: The primary objective of this study is to examine the performance of widely used sampling variance estimators in international large-scale assessments (ILSAs): *Balanced Repeated Replication* (BRR) and *Jackknife Repeated Replication* (*JK2*; *Half* and *Full* variants). We also investigate the impact of *Fay* modification factors (*10%*, *30%*, and *50%*) and evaluate *Bootstrapping* as an alternative variance estimator for clustered sampling.

Methods: A Monte Carlo simulation mirroring the sampling design of a TIMSS Grade 4 population was conducted. Probability samples were repeatedly drawn under a two-stage stratified cluster sampling design with schools as primary sampling units and classes as secondary sampling units. The simulation varied school sample sizes (30, 50, 100, and 150 schools), considered scenarios with odd numbers of schools within explicit strata, and examined different treatments of school non-response. For each scenario, 1,000 samples were drawn, and the performance of the variance estimators was assessed by comparing the estimated sampling errors with the "true" sampling errors approximated from the sampling distributions of the estimates.

Results: The results showed that, for smooth statistics such as means, *BRR without Fay adjustment*, *JK2*, and *Bootstrapping* produced highly accurate variance estimates, with *JK2* showing the highest precision across conditions. For non-smooth statistics, including percentiles, *Bootstrapping* and *BRR* outperformed *JK2* in both accuracy and precision. The *JK2 Half* and *Full* variants performed virtually identically, suggesting no precision advantage of the computationally more intensive *Full* variant. *Fay* adjustments did not improve *JK2* performance, whereas *BRR* performed better with lower *Fay* factors (10% and 30%) than with the conventional 50% adjustment. Under school non-response scenarios, common approaches to variance strata construction led to systematic over- or underestimation of sampling variance.

Conclusions: Overall, the study provides empirical guidance on the application of replication-based variance estimators in ILSAs under varying statistical conditions and sampling designs.

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

Introduction

International large-scale assessments in education (ILSAs) 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 (Author et al., 2021).

The sampling process in ILSAs is complex and can be summarized as a two-stage stratified cluster sampling design (Author et al., 2021). In the first stage, schools—serving as primary sampling units (PSUs)—are sampled with probabilities proportional to size (PPS) within explicit strata (Snijders & Bosker, 2011). In the second stage, students or classrooms are typically selected using systematic random sampling (SRS) (see class sampling in TIMSS [Siegel & Foy, 2024] and student sampling in PISA [OECD, 2024]). During data collection, sampling units at both stages are subject to non-response, that is, schools and students may decide not to participate in the survey (Author, 2024). In ILSAs, the combination of complex sample design, weighting adjustments for nonresponding units, and potential poststratification procedures directly impacts the level of error in statistical estimates (Kish, 1965; Mang et al., 2024). Given this complexity, accurate and precise sampling variance estimation is crucial for reliable statistical inference, as it is inappropriate to apply "textbook" variance estimators that assume simple random sampling under multistage stratified cluster designs (Rust & Rao, 1996; Rutkowski et al., 2010).

Several approaches have been proposed to address variance estimation in complex sample surveys, including linearization techniques (Deville, 1999). However, replication methods—also referred to as resampling techniques—are among the most widely used approaches in survey methodology (Valliant et al., 2018). Although *Bootstrapping* is popular in survey and non-survey statistics more broadly (Valliant et al., 2018), *Balanced Repeated Replication* (BRR) and paired *Jackknife Repeated Replication* (JK2) are the most frequently employed methods for estimating sampling variance in ILSAs (Author, 2020). Variance estimators vary both in methodology and in their application across studies, for example *Fay's BRR* in PISA (OECD, 2024), *JK2 Half* in ICILS (Schulz, 2020), and *JK2 Full* 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 underscores the importance of investigating this issue to enhance understanding of and improve ILSA survey methodology.

By focusing on two widely used estimators—*BRR* and *JK2*—the central aim of this 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.

Theoretical Background

Variance Estimation in ILSAs

In ILSAs, uncertainty in the estimation of population parameters arising from the complex sampling design is commonly quantified in terms of sampling error[1], defined as the square root of the sampling variance. (Kish, 1965). To estimate sampling variance under the complex multistage sampling designs used in ILSAs, replication-based methods are widely employed (Rust & Rao, 1996; Valliant et al., 2018). Based on the systematic sampling order, sampled schools serving as PSUs are paired into variance strata for variance estimation (Gonzalez & Foy, 1996; Author et al., 2021). Using these variance strata, replicate weights are generated and sampling variance is estimated from the variability of replicate estimates around the full-sample estimate (Judkins, 1990; Choudhry & Valliant, 2004). The sampling error SE_ε is generally computed as

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

1 In ILSAs, achievement is reported using plausible values, and reported standard errors therefore combine sampling and imputation error (Mislevy et al., 1992; Rutkowski et al., 2010). Sampling variance is estimated separately for each plausible value and averaged to obtain the final sampling error (Gonzalez, 2024); this study only focuses on the sampling component of uncertainty.

where ε_r denotes the statistic of interest estimated using replicate weight r, ε_0 the same statistic estimated using the total sampling weight, R as the total number of replicate weights, and k as a scaling factor conditional on the specific replication method employed in the ILSA (Gonzalez, 2024). While replication methods vary in implementation, they share this common principle and differ primarily in how replicate weights and scaling factors are defined, including how variance strata are treated under non-response (Fishbein et al., 2024; OECD, 2024).

Replication Methods

To account for the uncertainty induced by complex sampling design, replication methods such as *BRR*, *JK2* and *Bootstrapping* have been developed (Bruch et al., 2011) [2]. 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. 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 assessments with a complex sampling design complex survey design, such as ILSAs.

In IEA studies, *JK2*[3] is the most widely applied approach for stratified multistage sampling designs (Wolter, 2007; Rust, 2013) and derives replicate weights by alternately deleting and inflating paired PSUs within variance strata. In each variance stratum, one PSU is retained with its weight doubled, $\boldsymbol{rw_1 = 2 \times w_1}$, while the other PSU is omitted by setting its replication weight to zero, $\boldsymbol{rw_2 = 0 \times w_2}$. Here, $\boldsymbol{w_i}$ denotes the estimation (full-sample) weight for PSU $\boldsymbol{i}$, and $\boldsymbol{rw_i}$ denotes its replication weight. For all PSUs in the remaining variance strata, the replication weight is unchanged, that is, $\boldsymbol{rw_i = w_i}$. The *Half* and *Full* variants differ in the number of replicate weights generated and in the associated scaling factor $\boldsymbol{k}$ (Author, 2020; Schulz, 2011, 2015; Fishbein et al., 2024), while using the same

[2] Detailed historical developments, mathematical formulations, and technical implementations of JK2, BRR (including Fay variants), and Bootstrapping—together with their use in survey methodology and ILSAs—are provided in Author et al. (2025).

[3] The original delete-one jackknife (*JK1*) is designed for single-stage samples and generates replicates by sequentially omitting individual observations. Because ILSAs employ stratified multistage cluster sampling, variance estimation is instead based on the paired *Jackknife* (*JK2*), which operates at the PSU level (Wolter, 2007).

sampling error equation (see *Equation 1*). In the *Half* method, the number of replicates equals the number of variance strata ($\boldsymbol{R = VS}$) and the scaling factor is $\boldsymbol{k = 1}$. In the *Full* method, the number of replicates is doubled ($\boldsymbol{R = 2 \times VS}$), and the scaling factor is $\boldsymbol{k = \frac{1}{2}}$.

BRR, used primarily in OECD studies, constructs replicate weights using balanced half-sample perturbations defined by a Hadamard matrix. The sampling error is computed using $Equation\ 1$, where the scaling factor is $k = \frac{1}{\mathrm{R}}$ and R equals the number of replicate weights (Judkins, 1990). To reduce instability arising from the exclusion of half the sample in each replicate, *BRR* is often implemented with a *Fay* modification, which perturbs weights by factors f and $2 - f$ and introduces an additional adjustment to the scaling factor (Fay, 1984; Dippo et al., 1984; Judkins, 1990). 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). An additional scaling factor $k = \frac{1}{R*(1-f)^2}$ was suggested to ensure "a reasonable estimate of the variance" (Judkins, 1990, p. 225).

While *Bootstrapping* is widely used in survey methodology more broadly (Efron, 1982; Rao & Wu, 1988; Beaumont & Émond, 2022), it is not commonly applied in ILSAs.

We summarize in **Table 1** the variance estimators that are used in major ILSAs.

[pls insert **Table 1** about here]

Differences in Building Variance Strata in ILSAs

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, when non-response occurs—as is common in ILSAs—differences in variance strata formation arise that can affect variance estimation. Because PSUs are subject to non-response, an important methodological question concerns how participating and non-participating PSUs are handled in the construction of variance strata.

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. Schools are sorted by the same sampling attributes used prior to sampling (e.g.,

school size and implicit stratification variables), and non-responding schools are excluded. Participating schools are then paired sequentially within each explicit stratum following the original sampling order. Because adjacent sampled schools are more likely to share similar characteristics, this re-pairing strategy reflects similarities induced by implicit stratification (Gonzalez & Foy, 1996).

In contrast, OECD's PISA (OECD, 2012; OECD, 2023) constructs variance strata at the time of sampling, prior to data collection. When non-response occurs, the original variance strata are retained rather than redefined. If a variance stratum originally contains two sampled schools and only one participates, replicate weights are generated solely from the participating school. This approach preserves the original sampling structure and avoids redefining variance strata based on participation status.

The key distinction between these approaches lies in whether variance strata are redefined after data collection (as in TIMSS, PIRLS, and TALIS) or retained from the original sampling design (as in PISA).

Odd number of PSUs in the explicit stratum

When an explicit stratum contains an odd number of schools, BRR and JK2 differ in how the unpaired PSU is handled. In JK2, the remaining unpaired school is subdivided into paired sub-PSU units (e.g., classes or students) so that each variance stratum contains two units (Fishbein et al., 2024). In BRR, the unpaired school is combined with two adjacent schools to form a triplet within the variance stratum (Mang et al., 2024). Replicate weights are then generated using structured weight perturbations derived from the Hadamard matrix (OECD, 2024).

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.

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). As highlighted by Valliant et al. (2018), replication methods 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. 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.

According to Valliant et al. (2018), 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. Simulation studies by Kovar (1985) and Judkins (1990) showed that *JK2* exhibits lower bias, greater stability, and more precise confidence intervals than *BRR* (with or without *Fay*) in high-variance populations. However, theoretical and empirical work has raised concerns about *JK2*'s performance for non-smooth statistics. Efron (1982) demonstrated that *JK1* produces unreliable variance estimates for statistics such as medians and quartiles, and Kovar, Rao, and Wu (1988) confirmed that the *JK2* performs worse than *BRR* for non-smooth statistics. They claim further that applying *Fay*'s *Method* to *BRR* consistently improves variance estimation for non-smooth statistics relative to both *JK2* and *standard BRR*.

A noted limitation of *BRR*, however, is its reliance on the *a priori* calculation and storage of replicate weights via the Hadamard matrix, whereas *JK2* weights can be generated systematically "on the fly." Another limitation of *BRR* is the occurrence of missing replicate estimates in small domains because each replicate effectively excludes half of the variance strata, increasing the likelihood that a domain is absent from a replicate. This issue is addressed by *Fay's Method*, which perturbs rather than completely deletes the weights, thereby reducing the occurrence of missing replicates. Judkins (1990) characterized *Fay's 50% Method* as a compromise between standard *BRR* and *JK2*, although a 30% Fay factor, representing a gentler perturbation of the weights, has been shown to provide more reliable estimates for quantiles. Judkins (1990) also demonstrated that *Fay's Method* is equivalent to *standard BRR* for linear estimators. By contrast, the application of Fay factors within the *JK2* family remains largely unexplored. Although *JK2* is substantially less prone than *BRR* to producing missing replicate estimates in small domains, such cases may still occur when the omitted variance stratum contains all observations for the domain of interest.

While *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.

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 1995 Technical Report (Gonzalez & Foy, 1996) and PISA Technical Report (OECD, 2002). Rust (2013) provides a useful theoretical overview of replication methods in ILSAs, but systematic empirical evaluations across methods and conditions remain scarce.

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 performance in estimating sampling variance for both smooth and non-smooth statistics, such as means, medians, quartiles, benchmark percentages, and correlations.

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 Accordingly, we investigate whether *Fay factors* of *10%* or *30%* improve performance relative to the commonly used *50% Fay factor*[4]. Although *Fay's Method* is typically applied to *BRR* to address missing replicates, this study also examines the effect of introducing *Fay factors* (*10%*, *30%*, and *50%*) to *JK2 Half*, particularly for non-smooth statistics.[5] In addition, we evaluate whether the increased computational effort required by *JK2 Full* yields gains in precision relative to *JK2 Half*.

[4] A Fay factor can take any value between 0 and 1.

[5] We also 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.

Our additional motivation in this research is to examine the effects of varying conditions of sample sizes and non-response. If there is *an odd number of PSUs* in a given explicit stratum, *BRR* addresses this by collapsing variance strata and forming triplets of PSUs. Based on Valliant et al. (2018), we hypothesize that such collapsing variance strata, or adopting a one-PSU-per-variance-stratum approach, may lead *BRR* to overestimate sampling variance, particularly in smaller samples. Conversely, Valliant et al. (2018) note that pairing sub-PSU units (e.g., classes or students) instead of PSUs may lead to underestimation of sampling variance; accordingly, we hypothesize that *JK2* may underestimate sampling variance in such scenarios.

As explained in the "Non-Response" section, one common approach is to retain PSUs in their original variance strata (*OrgStr*) or to re-pair only participating PSUs. Under the *OrgStr* approach, *BRR* treats each participating PSU as a single PSU within its original variance stratum, so that only one PSU contributes to variance estimation within that stratum. In contrast, under the standard BRR approach, participating PSUs are re-paired into new variance strata, ensuring that at least two PSUs contribute to the generation of replicate weights and, consequently, to the variance estimation.

In ILSAs, *JK2* has traditionally not been implemented using the *OrgStr* approach because the method is not well suited to variance strata containing a single PSU. For completeness, we nevertheless include *JK2 OrgStr* variants for both the Half and Full methods to enable direct comparisons with *BRR*. Following current ILSA practice, single-PSU variance strata are handled by subdividing the PSU into paired sub-PSU units, including orphan PSUs resulting from an odd number of schools within an explicit stratum. Because sub-PSU units within a school are generally more homogeneous than PSUs across schools, this approach may underestimate sampling variance, particularly under nonresponse when multiple single-PSU strata occur. Accordingly, results for the *JK2 OrgStr* variants should be interpreted with caution. The *JK2 OrgStr* implementation follows procedures used in TIMSS (Fishbein et al., 2024), whereas *BRR OrgStr* follows the approach used in PISA (OECD, 2024).

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?

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[6].

Simulated Population

We simulated a finite population of 5,000 schools and 500,857 students to analyze the performance of sampling variance estimators[7]. This population is designed to reflect the average population size observed in ILSAs, such as the TIMSS Grade 4 population (Siegel & Foy, 2024). 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. This results in a symmetrical distribution of student enrollment with an average school having around 100 students. We summarized simulated population characteristics in **Table 2.**

[pls insert **Table 2** about here]

[6] 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.
[7] The simulated population minimizes certainty schools, avoiding repeated selection across samples and obviating the need for a finite population correction.

Two achievement scores were generated with intraclass correlation coefficients (ICCs) of 25% and 50% to represent average and high levels of school-level clustering[8]. Achievement was simulated using a linear model with school-level covariates and random school effects, then rescaled to the TIMSS metric (mean = 500, SD = 80). Students were randomly assigned to classrooms (average class size = 20), resulting in 1–10 classrooms per school. A binary gender variable was also generated.

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 in **Table 3.**

[pls insert **Table 3** about here]

At the first stage, schools were selected using systematic PPS sampling with implicit stratification by school size and explicit stratification by school type, following the standard TIMSS procedures (Siegel & Foy, 2024). School sample sizes ranged from 30 to 150 schools.

To evaluate estimator performance under realistic field conditions, simulations were conducted with and without school nonresponse. In the nonresponse condition, approximately 50% of sampled private schools did not participate, resulting in an overall school nonresponse rate of 12%–13%. School weights, including nonresponse adjustments, were calculated according to the standard TIMSS procedures (Siegel & Foy, 2024).

At the second stage, one or two classes were sampled within each selected school, following the standard TIMSS design (Siegel & Foy, 2024). Final student weights were constructed using the corresponding school and class weights, with no adjustment for student nonresponse because student nonresponse was not simulated.

Variance Estimators and Replicate Weights

We evaluated the performance variance estimators categorized into three main replication methods: *JK2*, *BRR*, and *Bootstrapping*. For the *Jackknife* method, we analyzed *JK2* in both *Full* and *Half* variants. For *BRR*, we examined *BRR* without a *Fay*

[8] School-level ICC estimates for TIMSS 2023 Grade 4 mathematics are presented in **Appendix Tables Table Appendix A 1**.

factor as well as Fay-adjusted *BRR* variants. Additionally, we explored the *Fay* factors of 10% ($f = 0.1$), 30% $f = 0.3$, and 50% $f = 0.5$ for *JK2 Half* and *BRR* estimators. If a *Fay* factor is applied to *JK2 Half*, the MSE is scaled by $k = \frac{1}{(1-f)^2}$, whereas for BRR the corresponding scaling factor is $k = \frac{1}{R*(1-f)^2}$. In addition, we examined variants that retain the original variance strata structure (*OrgStr*) under non-response, including *BRR OrgStr*, *JK2 Half OrgStr*, and *JK2 Full OrgStr*. For *BRR-based* methods, Hadamard matrices were constructed with dimension H equal to the smallest multiple of four greater than or equal to the number of sampled schools S.

Also, we implemented a bootstrap variance estimator tailored to the clustered sampling structure of ILSAs. The bootstrap procedure uses the same variance strata as the *JK2* design, treating paired schools as PSUs, and generates replicate weights by randomly inflating one PSU and deflating the other within each variance stratum.[9].

The **Table 4** summarizes the variance estimators examined in this study, the corresponding number of originally sampled schools (S), variance strata (VS), the number of replicate weights (R), and the number of participating schools (P). As shown, the number of replicate weights varies by method—which coincides with the number of variance strata for *JK2 Half* and *BRR* (including *Fay* variants), and which coincides with the number of sampled or participating schools for *JK2 Full* and *Bootstrapping*.

[pls insert **Table 4** about here]

Estimated Parameters of Interest

The Horvitz–Thompson (HT) estimator was used to estimate population parameters of interest (Horvitz & Thompson, 1952; Mang et al., 2021). The analyses included both smooth and non-smooth statistics, comprising means, differences in means, percentiles, benchmark percentages, and Pearson correlations. Detailed results for each parameter are presented in the corresponding tables.

[9] A full technical description of the bootstrap implementation for ILSAs—including the resampling scheme, replication factors, and mathematical formulation—is provided in the IEA R&D report (Author et al., 2025; see also Beaumont & Émond, 2022).

JK2 and BRR were implemented in R (R Core Team, 2022) using the IEA IDB Analyzer (IEA, 2025), whereas bootstrapping was implemented in R using the survey package (Lumley, 2004). Additional implementation details are available in the respective software documentation[10].

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.

$$se(\hat{\epsilon}_i)^t = \sqrt{var(\hat{\epsilon}_i)^t} \tag{2}$$

The population parameter of interest is denoted by $\hat{\epsilon}_i$ and its sampling variance is denoted by $var(\hat{\epsilon}_i)^t$. t represents the respective variance estimator 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. The following equation indicates the average sampling error:

$$\overline{se}(\hat{\epsilon})^t = \frac{\sum_{i=1}^{1000} se(\hat{\epsilon}_i)^t}{1000} \tag{3}$$

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 (Author et al., 2021).:

$$se(\hat{\epsilon})^* = \sqrt{\frac{\sum_{i=1}^{1000}(\hat{\epsilon}_i - \hat{\epsilon}^*)^2}{1000 - 1}} \tag{4}$$

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

[10] While *JK2*, *BRR*, and *Fay-adjusted BRR* are implemented in standard software used for ILSA analyses—including the IEA IDB Analyzer, RALSA, EdSurvey, and REPEST—bootstrap variance estimation is currently less widely supported and is typically implemented using general-purpose survey software in R, such as the *survey* and *svrep* packages (Lumley, 2004; Schneider, 2023).

The *Relative Bias* is defined by 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:

$$Relative\ Bias_{\overline{se}(\hat{\epsilon})^t} = \frac{\overline{se}(\hat{\epsilon})^t}{se(\hat{\epsilon})^*} \tag{5}$$

The *Stability* of the sampling error is defined by 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:

$$Stability_{\overline{se}(\hat{\epsilon})^t} = \frac{sd\,(se(\hat{\epsilon})^t)}{se(\hat{\epsilon})^*} \tag{6}$$

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). To enable meaningful comparisons across variance estimators and statistics, performance is evaluated using *Relative Bias* and *Stability*. *Relative Bias* measures the ratio of the estimated sampling error to the true sampling error, with values above one indicating overestimation and values below one indicating underestimation. *Stability* captures the variability of the estimated sampling error across repeated samples and can be interpreted as the standard error of the sampling error; lower values indicate a more precise and stable variance estimator.

Results

The results across all tables in this section summarize the performance of variance estimators for the statistical parameters analyzed, focusing on comparisons of *Relative Bias* and *Stability*. Results were averaged across the two achievement-score conditions (ICC = 25% and 50%) to provide an overall summary of estimator performance across moderate and high levels of school clustering. Results for each ICC condition are reported separately in the supplementary documents. The results are reported for four sample size scenarios (very small, small, medium, and large) under three conditions: *Regular Sample*, *Odd-Number of Schools*, and *Non-Response*. Within each results table, we highlight the estimators closest to a *Relative Bias* of 1 and those closest to a *Stability* value of 0.

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.

[pls insert **Table 5** about here]

In the ***Regular Sample Scenario***, all estimators show minimal *Relative Bias*, performing similarly even in very small samples. *Bootstrapping* yields the lowest *Relative Bias* in medium samples, whereas *JK2* estimators consistently provide the best *Stability*. *BRR* variants and *JK2* variants perform identically within their respective families. Across estimators, *Stability* improves as sample size increases.

In the ***Odd Number of Schools Scenario***, overall patterns remain similar. *JK2* and *standard BRR* perform best in very small samples, while *Bootstrapping* shows slightly lower *Relative Bias* in small and large samples. *BRR with a 50% Fay factor* performs worse in smaller samples, whereas lower *Fay factors* (*10%* and *30%)* produce results closer to *standard BRR*. *JK2* estimators again exhibit the highest *Stability* across all sample sizes.

In the ***Non-Response Scenario***, *Relative Bias* decreases as sample size increases for most estimators. *Standard BRR* and *JK2* estimators show similar *Relative Bias* patterns, while *Bootstrapping* performs slightly better in terms of bias but with weaker *Stability*. *BRR Fay 50% OrgStr* strongly overestimates sampling error, whereas *JK2 OrgStr* variants substantially underestimate it despite showing high *Stability*. No differences are observed between *JK2 Full* and *Half* estimators or between *Fay-adjusted* and standard *JK2* variants.

Sub-Group Means (Gender = Girls)

Table 6 presents the variance estimator comparison for mean estimates of girls comprising a sub-group of interest.[11]

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

Overall, the results closely mirror those observed for overall means. Across all sampling scenarios, *JK2* estimators consistently achieve the highest *Stability*, while *BRR* and its *Fay-adjusted* variants show comparable *Relative Bias*, except for *BRR with a 50%*

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

Fay factor, which overestimates sampling error in small samples. *Bootstrapping* remains competitive in terms of *Relative Bias* but exhibits weaker *Stability*, particularly for smaller sample sizes. As expected, sub-group analyses yield slightly higher variability than overall estimates, but no estimator displays qualitatively different behavior. These findings indicate that conclusions drawn for overall means generalize directly to sub-group mean estimates.

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]

In the ***Regular Sample Scenario***, all variance estimators produce *Relative Bias* values close to 1, indicating high accuracy across sample sizes. *Bootstrapping* shows slightly lower *Relative Bias* in very small samples but remains comparable to other estimators overall. As in the previous results for means, *JK2* estimators provide the best *Stability*, while *Bootstrapping* shows the weakest *Stability*. Precision improves for all estimators as sample size increases.

In the ***Odd Number of Schools Scenario***, the overall pattern remains similar. *Standard BRR* and *JK2* estimators maintain *Relative Bias* values close to 1 across sample sizes, while *Bootstrapping* slightly underestimates sampling variance in smaller samples but aligns with other estimators in larger samples. Stability patterns again favor *JK2* estimators, which consistently achieve the highest precision.

In the ***Non-Response Scenario***, *Relative Bias* shows slightly greater variation, with most estimators exhibiting mild underestimation in smaller samples. *Bootstrapping* performs similarly but with somewhat weaker *Stability*. *JK2 OrgStr* variants notably underestimate sampling variance despite showing the strongest *Stability*. Overall, *Stability* decreases as sample size increases, and performance patterns closely mirror those observed for overall mean estimates.

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

Table 8, **Table 9**, and **Table 10** indicate *Relative Bias* and *Stability* values across percentiles, 25th, 50th (median), and 75th percentiles at the overall level.[12]

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

In the ***Regular Sample Scenario***, *standard BRR* and *Bootstrapping* generally produce the lowest *Relative Bias* across percentiles, with *Relative Bias* decreasing as sample size increases. *Bootstrapping* performs similarly in terms of bias but shows weaker *Stability* in very small samples. *JK2* estimators exhibit larger and less consistent *Relative Bias* across percentiles and sample sizes. Among *BRR* variants, lower *Fay* factors (*10%* and *30%*) perform better than the *50%* adjustment. Overall, *standard BRR* achieves the best *Stability*.

In the ***Odd Number of Schools Scenario***, *standard BRR* and *Bootstrapping* continue to perform well, maintaining low *Relative Bias* across percentiles. *Bootstrapping* shows slightly weaker *Stability* in smaller samples. *Fay-adjusted BRR* variants remain comparable at lower adjustment levels but perform worse at the *50%* level. *JK2* estimators display consistently higher *Relative Bias* and weaker *Stability* relative to *BRR-based* methods.

In the ***Non-Response Scenario***, all estimators' performance deteriorates somewhat due to increased design complexity. *Standard BRR* and *Bootstrapping* again show the most favorable balance of *Relative Bias* and *Stability*, particularly in larger samples. *BRR Fay 50% OrgStr* substantially overestimates sampling variance, whereas *JK2 OrgStr* variants strongly underestimate it despite showing comparatively high *Stability*. Overall, Bootstrapping remains competitive, particularly in medium-sized samples.

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**.[13]

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

In the ***Regular Sample Scenario***, differences among variance estimators are minimal. All methods produce *Relative Bias* values close to 1 across sample sizes,

[12] We also analyzed the percentiles 25th, 50th, and 75th by sub-group levels ($gender = girl$). The tables for these analyses are provided in the **Error! Reference source not found.**, 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.

[13] 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.

indicating accurate variance estimation. *Standard BRR*, *Fay-adjusted BRR* variants, *JK2* estimators, and *Bootstrapping* show comparable performance, with only minor differences in *Stability*. *Bootstrapping* tends to exhibit slightly weaker *Stability* in very small samples, while *JK2* estimators generally achieve slightly better precision.

In the ***Odd Number of Schools Scenario***, results closely resemble those observed under regular sampling conditions. *BRR*, its *Fay-adjusted* variants, and *JK2* estimators maintain *Relative Bias* values near 1 and low *Stability* values across sample sizes. *BRR with a 50% Fay adjustment* performs slightly worse in smaller samples, while lower *Fay* factors remain closer to *standard BRR* performance. Overall differences among estimators remain small.

In the ***Non-Response Scenario***, performance differences become somewhat more pronounced. *BRR* and *JK2* estimators continue to produce *Relative Bias* values close to 1, whereas *Bootstrapping* shows slightly lower *Relative Bias* and somewhat weaker *Stability* in smaller samples. *JK2 OrgStr* variants tend to underestimate sampling variance, while *BRR OrgStr* shows modest overestimation. Despite these differences, most estimators continue to perform similarly in estimating benchmark percentages.

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[14]; 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[15].

[14] Because correlation statistics are mainly relevant for secondary analyses, detailed results are reported in Appendix Table C.1.
[15] We acknowledge the limits of such approach, such as equal contribution of all scenarios/analysis types to the average.

Table 14 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 14**, **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.

Regular Sample Scenario

In the ***Regular Sample Scenario***, according to **Table 14** and **Figure 1**, *Bootstrapping* shows the lowest *Relative Bias* across most sample sizes. *BRR* and its *Fay variants* also perform well, with only slightly higher bias but strong *Stability*. *JK2* estimators show comparable accuracy but somewhat weaker *Stability*, particularly in smaller samples. Across all estimators, *Stability* improves as sample size increases.

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 14** 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.

Non-Response Scenario

As indicated in **Table 14** 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*.

Summary

Across all scenarios, the following patterns are evident:

i. As one would expect, 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.

ii. *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.

iii. *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*.

iv. *JK2* estimators perform well closely following *BRR* and *Bootstrapping* in both *Relative Bia*s. However, their precision is more volatile , especially in smaller sample size*s*. 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*.

Discussion and Recommendations

For practical applications in ILSAs, it is practical 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***.

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 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*?

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.

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 more comparable results 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.

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. This effect would be exacerbated with higher numbers of plausible values, as used in PISA and recommended by Jewsbury et al. (2025). 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.

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.

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 further 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 sampling 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.

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*[16] 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 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.

Conclusions and Limitations

This study makes an important contribution to the understanding of sampling variance estimators in the context of ILSAs. Understanding the strengths and weaknesses of variance estimation techniques are 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 accuracy and 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, leading to incorrect conclusions about group differences. Such errors in inference are not limited to subgroup comparisons but extend to other statistical comparisons, such as trend analyses within a country over multiple years.

[16] www.iea.nl

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 or occurring in a systematic manner, 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 first step for further exploration, offering insights and benchmarks for future studies to refine and improve variance estimation in ILSAs.

References

Author et al. (2013, 2020, 2021, 2024, 2025).

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.

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.

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.

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)

Jewsbury, P.A., McCaffrey, D.F., Jia, Y. and Gonzalez, E.J. (2025), How Many Plausible Values?. Journal of Educational Measurement, 62: 531-558. https://doi.org/10.1111/jedm.70000

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

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

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 (2002), PISA 2000 technical report. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/publications/technical-report/PISA%202000%20Technical%20report.pdf

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., & 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

Tables

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	$\frac{1}{R*(1-f)^2}$	80	80
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

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 25^{th}, 50^{th}, 75^{th} 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 25^{th}, 50^{th}, 75^{th} 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

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

Table 4: Variance Estimators Examined in the Study

Variance Estimator	Number of PSUs S or P	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: S denotes sampled schools; P denotes participating schools. When there is 100% response rate simulated, $S = P$.

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

	Relative Bias				Stability			
Variance Estimator Method	**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

	Relative Bias				Stability			
Variance Estimator Method	**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 - 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, 25^{th,} 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.

Figures

Figure 1

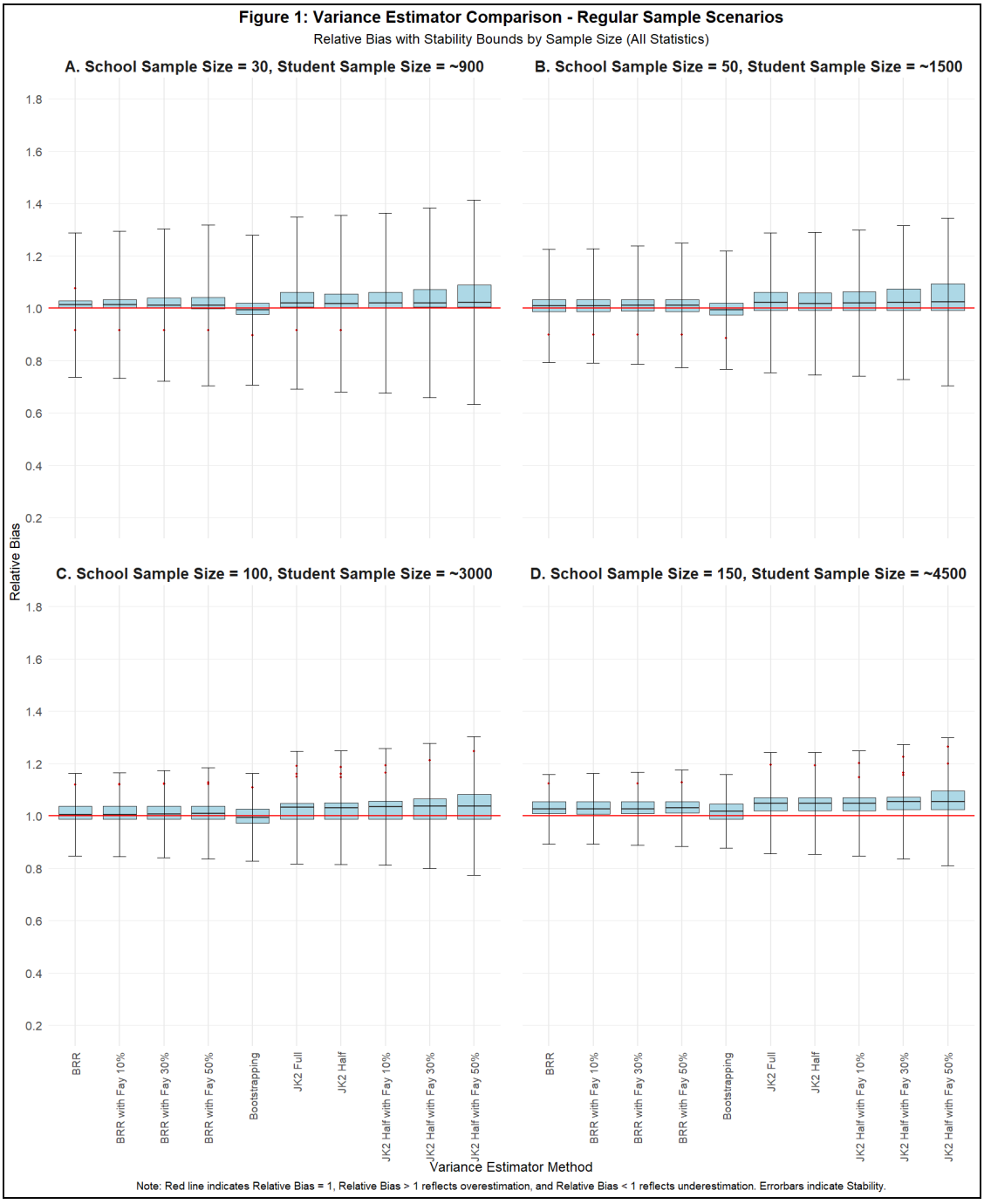


Figure 2

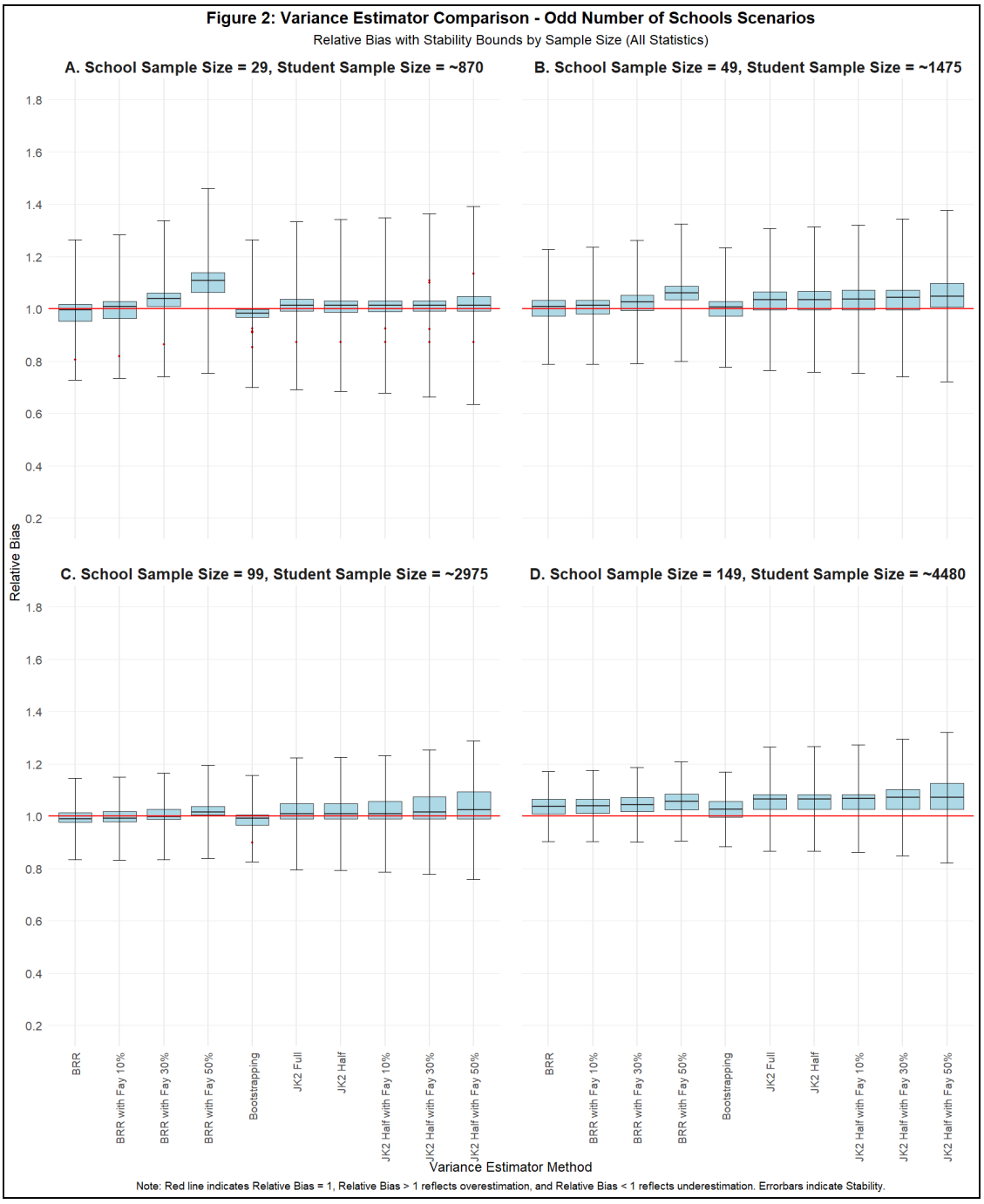
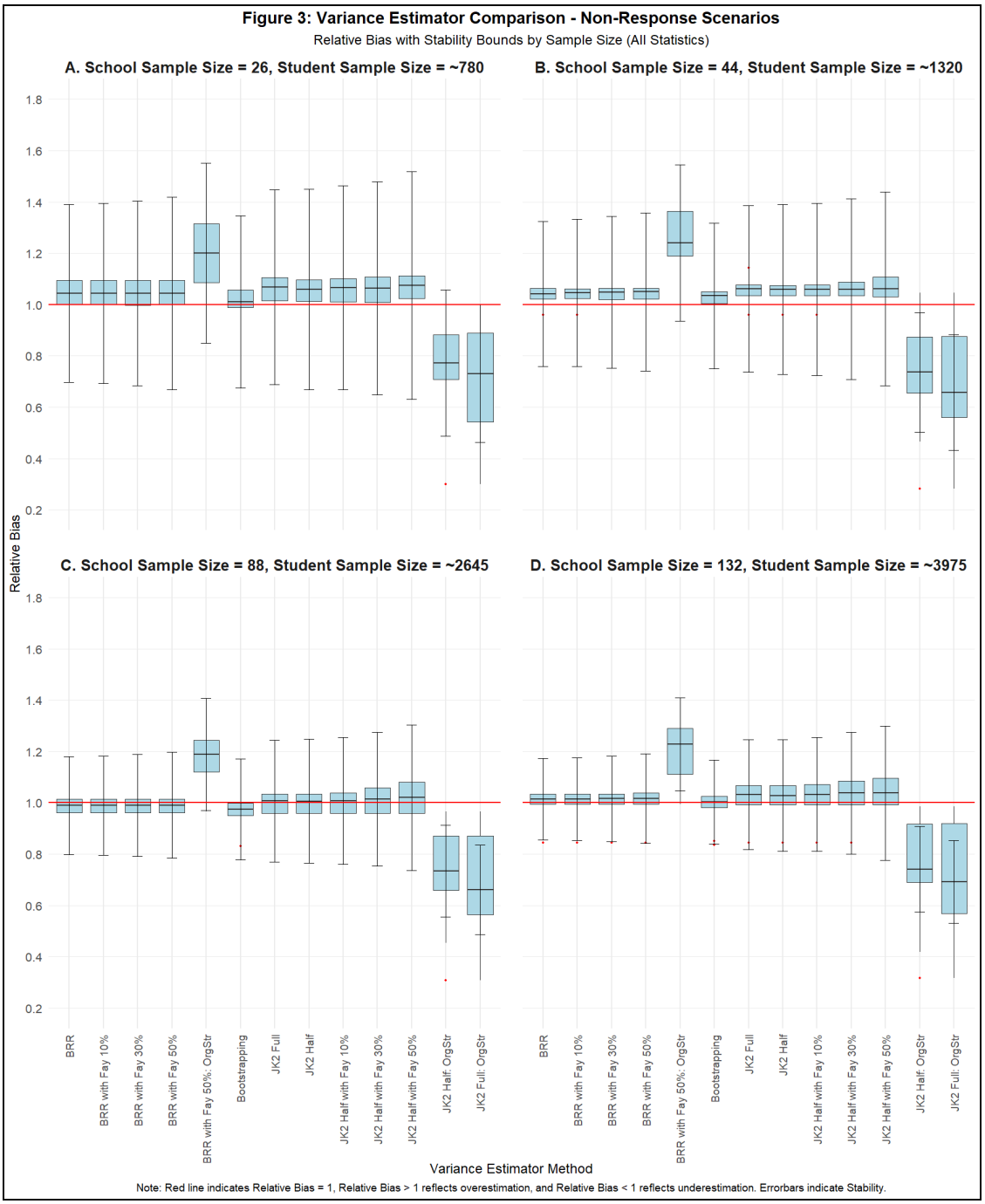

Figure 2: Variance Estimator Comparison - Odd Number of Schools Scenarios
Relative Bias with Stability Bounds by Sample Size (All Statistics)
A. School Sample Size = 29, Student Sample Size = ~870
B. School Sample Size = 49, Student Sample Size = ~1475
C. School Sample Size = 99, Student Sample Size = ~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.

Appendix Tables

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

Appendix Tables

Table Appendix A 1: Mathematics Achievement Statistics - TIMSS 2023

Country	Mean	Standard Deviation (SD)	Intra-Class Coefficient (ICC)
Spain	504	75	0.18
Sweden	534	75	0.15
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.

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

	Relative Bias				Stability			
Variance Estimator Method	**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.

Appendix C 1: Variance Estimator Comparison - Pearson Correlation between Score 1 and Score 2 (Overall)

Relative Bias and Stability Across Sampling Scenarios

	Relative Bias				Stability			
Variance Estimator Method	**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.