

Design-Consistent Variance Estimation in Multistage Complex Surveys: A Simulation and MICS-Based Comparative Study

Ali Satty^{1,*}, Zakariya M. S. Mohammed^{1,2}

¹Department of Mathematics, College of Science, Northern Border University, Saudi Arabia

²Center for Scientific Research and Entrepreneurship, Northern Border University, Saudi Arabia

Received October 30, 2025; Revised December 30, 2025; Accepted January 15, 2026

Cite This Paper in the Following Citation Styles

(a): [1] Ali Satty, Zakariya M. S. Mohammed, "Design-Consistent Variance Estimation in Multistage Complex Surveys: A Simulation and MICS-Based Comparative Study," *Mathematics and Statistics*, Vol.14, No.1, pp. 91-97, 2026. DOI: 10.13189/ms.2026.140108.

(b): Ali Satty, Zakariya M. S. Mohammed (2026). *Design-Consistent Variance Estimation in Multistage Complex Surveys: A Simulation and MICS-Based Comparative Study*. *Mathematics and Statistics*, 14(1), 91-97. DOI: 10.13189/ms.2026.140108.

Copyright ©2026 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Ignoring survey design features such as clustering, stratification, and unequal weighting can lead to underestimated standard errors (SEs) and misleading inference in regression models. This study compares three design-consistent variance estimators, Taylor linearization (TL), Fay's balanced repeated replication (BRR), and the Rao–Wu–Yue bootstrap, using Monte Carlo simulations based on the two-stage stratified structure of UNICEF's Multiple Indicator Cluster Surveys (MICS). Nine scenarios combine intra-class correlation ($ICC = 0.01-0.10$) with weight variability ($CV_w = 0.2-1.0$) to assess 95% coverage, SE calibration, and confidence interval (CI) width. Coverage was generally near nominal when clustering was weak to moderate ($ICC \leq 0.05$), with mild under-coverage (about 90%) at $ICC = 0.10$ across methods. SEs were well calibrated (SE-ratios $\approx 0.93-1.03$). CI width was driven primarily by weight heterogeneity, increasing markedly with larger CV_w , whereas ICC had a smaller impact. In an application to 2018–2019 MICS data on childhood diarrhea, point estimates (odds ratios) were identical across methods; BRR and RWY bootstrap yielded slightly wider, more conservative CIs. Overall, TL is most efficient under moderate design effects, while replication methods offer greater robustness when clustering and weight dispersion are high, providing practical guidance for MICS-type analyses.

Keywords Complex Survey Design, Variance Estimation, Taylor Linearization, Balanced Repeated Replication, Rao–Wu–Yue Bootstrap

1 Introduction

Accurate estimation of standard errors (SEs) is fundamental for valid statistical inference in survey-based regression models. In complex survey designs, naïve analyses that ignore design features such as clustering, stratification, and unequal weighting can yield biased SEs, misleading confidence intervals (CIs), and invalid hypothesis tests [1, 2, 3]. Underestimated variances lead to overstated statistical significance, while overestimated ones reduce efficiency and statistical power. These issues are particularly critical in household and health surveys such as UNICEF's Multiple Indicator Cluster Surveys (MICS), where results inform national policy and monitoring of Sustainable Development Goals (SDGs). Ensuring accurate and design-consistent SEs is therefore not merely a technical concern but essential for evidence-based decision-making in public health and social research.

Complex multistage surveys commonly feature stratification, clustering, and unequal selection probabilities, all of which affect sampling variability and require specialized variance estimation methods [1]. In typical two-stage designs, PSUs are selected within strata and households within PSUs, creating intra-class correlation (ICC), while sampling weights adjust for unequal selection and nonresponse. The MICS program follows this structure across more than 100 countries, combining regional and urban–rural stratification with clustered sampling. Consequently, accurate analysis of MICS data must account for clustering and weight variability (CV_w). Approaches such as Taylor linearization (TL) [3], balanced repeated replication (BRR) [4], and the bootstrap [1, 5, 6] are widely applied [2, 7, 8], yet their comparative performance under realistic MICS-like conditions remains insufficiently ex-

plored.

Although each design-consistent variance estimator has well-documented theoretical properties, comparative empirical evidence across realistic survey configurations remains limited. Prior studies often focus on single estimators or idealized designs, rarely exploring the joint effects of ICC and weight heterogeneity on estimator performance [4, 6, 9]. As a result, applied researchers lack practical guidance on which method performs best across varying design complexities common in international surveys such as MICS, Demographic and Health Surveys (DHS), and Living Standards Measurement Study (LSMS). Furthermore, most existing comparisons emphasize asymptotic properties, while simulation-based evaluations under finite-sample, multistage conditions mirroring real survey designs are scarce. This gap is especially important because survey analysts frequently encounter moderate clustering, variable PSU sizes, and high weight dispersion, all of which challenge classical linearization formulas and motivate replication-based alternatives [10, 11].

The present study aims to address this gap by providing a systematic comparison of three design-consistent variance estimators, TL, Fay's BRR, and the bootstrap, within MICS-like two-stage survey structures. Using an extensive simulation framework, the study evaluates their performance across nine scenarios combining different levels of ICC and CV_w . The comparison focuses on key performance metrics, including 95% coverage probability, SE calibration, and CI width, which jointly assess the accuracy, reliability, and efficiency of the estimators. Complementing the simulation, a MICS application is used to illustrate practical implications for real-world data analysis. The study contributes by (i) clarifying the conditions under which each estimator performs optimally, (ii) quantifying trade-offs between efficiency and robustness, and (iii) offering practical recommendations for analysts selecting design-consistent variance estimators in complex survey regression modeling. Together, these contributions advance both the methodological understanding and applied practice of variance estimation in modern household surveys.

2 Materials and Methods

Let $\hat{\theta}$ denote an estimator of a population parameter θ (e.g., a regression coefficient) computed from a complex survey with stratification, clustering, and unequal weights. The goal is to estimate its sampling variance

$$V_p(\hat{\theta}) = E_p \left[(\hat{\theta} - \theta)^2 \right], \quad (1)$$

where $E_p(\cdot)$ is the expectation under the sampling design p . Because $V_p(\hat{\theta})$ depends on the multistage selection mechanism and is analytically intractable for nonlinear estimators, it is approximated using the design-consistent estimators introduced above, Taylor linearization (TL), Fay's BRR, and the Rao–Wu–Yue bootstrap. All are asymptotically design-consistent under mild regularity conditions, but they differ in how they approximate the sampling distribution and in their robustness to clustering and high weight dispersion [3, 4, 5].

2.1 Taylor Linearization (TL)

TL approximates a nonlinear statistic using a first-order Taylor expansion around the population mean [2, 3]. The method relies on the idea that many estimators, such as ratios or regression coefficients, behave almost linearly under small perturbations. By quantifying how the estimator changes when a PSU or weight is slightly modified, TL constructs a linearized variable that captures this local sensitivity. This enables nonlinear estimators to be treated with variance formulas similar to those for sample means, making TL highly efficient for large, stratified surveys.

Formally, for a smooth function $f(\mathbf{y})$ of survey variables \mathbf{y} ,

$$f(\bar{\mathbf{y}}) \approx f(\mu_{\mathbf{y}}) + \nabla f(\mu_{\mathbf{y}})^\top (\bar{\mathbf{y}} - \mu_{\mathbf{y}}), \quad (2)$$

where $\nabla f(\mu_{\mathbf{y}})$ is the gradient vector. This approximation expresses the nonlinear estimator in terms of a locally linear function of the data, allowing its variability to be assessed using differences between PSU-level means of the linearized variable.

Let w_{hij} denote the final analysis weight for observation i in PSU j and stratum h . Define the stratum- and PSU-level means of the linearized variable z_{hij} by

$$\bar{z}_h = \frac{1}{n_h} \sum_j z_{hj}, \quad \bar{z}_{hj} = \frac{1}{m_{hj}} \sum_{i=1}^{m_{hj}} z_{hij}, \quad (3)$$

where H is the number of strata, n_h the number of PSUs in stratum h , and m_{hj} the number of sampled households in PSU j of stratum h . The TL variance estimator for $\hat{\theta}$ is

$$\widehat{V}_{\text{TL}}(\hat{\theta}) = \sum_h 1^H \frac{1}{n_h(n_h - 1)} \sum_{j=1}^{n_h} (\bar{z}_{hj} - \bar{z}_h)^2. \quad (4)$$

From a practical perspective, TL performs very well when the survey weights do not vary too drastically across observations and when within-PSU homogeneity is moderate. However, when ICC or CV_w is high, the neglected higher-order terms may lead to mild underestimation of variance [12, 13]. This motivates alternative methods such as survey bootstrapping when complex nonlinear behavior or strong clustering is present.

2.2 Balanced Repeated Replication (BRR)

BRR constructs a set of half-sample replicates by systematically selecting and perturbing PSUs using a Hadamard matrix [4]. Let R denote the number of replicates and $\hat{\theta}^{(r)}$ the replicate estimate. The BRR variance estimator is

$$\widehat{V}_{\text{BRR}}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta})^2. \quad (5)$$

Conceptually, BRR estimates variance by repeatedly recomputing the statistic on systematically perturbed half-samples, quantifying how sensitive the estimator is to balanced PSU perturbations. Each replicate's weights are defined as

$$w_{hij}^{(r)} = w_{hij} \left(1 + \delta_{hj}^{(r)} \right), \quad (6)$$

where $\delta_{hj}^{(r)} \in \{-1, +1\}$ indicates which PSU is up- or down-weighted in replicate r .

Fay’s BRR modification introduces a shrinkage factor ρ that reduces the perturbation of replicate weights:

$$w_{hij}^{(r)} = w_{hij} \left[1 + (1 - \rho) \delta_{hj}^{(r)} \right]. \tag{7}$$

When $\rho = 0$, Fay’s BRR reduces to classical BRR. A value of $\rho = 0.5$ is commonly used for moderate stabilization [4, 10]. Fay’s BRR preserves the BRR design but produces smoother, more stable variance estimates in small or unbalanced strata.

2.3 Bootstrap for Complex Surveys

The Rao–Wu–Yue (RWY) bootstrap [5, 6, 14, 15] resamples PSUs to mimic the variability generated by the actual two-stage sampling design. Unlike TL, which relies on derivative-based approximations, the bootstrap creates many synthetic surveys drawn under the same stratified cluster structure. The variability across these replicates provides a direct estimate of sampling variance. This makes the bootstrap particularly useful for complex or highly nonlinear estimators, where first-order approximations may be inadequate.

For bootstrap replicate r : (i) resample n_h PSUs with replacement from each stratum h ; (ii) compute replicate weights $w_{hij}^{(r)}$ by scaling the original weights of the selected PSUs; and (iii) re-estimate the statistic $\hat{\theta}^{(r)}$ using the replicate design.

The bootstrap variance estimator is

$$\widehat{V}\text{Boot}(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^R \left(\hat{\theta}^{(r)} - \bar{\theta}\text{Boot} \right)^2, \tag{8}$$

where $\bar{\theta}\text{Boot} = \frac{1}{R} \sum_{r=1}^R \hat{\theta}^{(r)}$. This replication-based approach closely reproduces the finite-sample distribution of the estimator, making it robust against strong clustering, large weight variation, or highly nonlinear estimators such as odds ratios or machine-learning predictions. In practice, using around $R = 80$ bootstrap replicates provides stable results for most MICS/DHS-style surveys [5, 11]. The theoretical justification for such resampling in complex, dependent designs is well established in [16, 17], supporting its asymptotic validity.

3 Simulation Study

3.1 Design and Data Generating Mechanism

To evaluate the performance of the variance estimators, a Monte Carlo simulation was designed to reflect the two-stage stratified sampling structure of the MICS survey. The simulated population comprised eight strata, each containing 30 primary sampling units (PSUs), for a total of 240 PSUs. Within each PSU, 30 households were sampled at the second stage. For each PSU, a finite household frame six times larger than the number sampled was created to ensure realistic within-PSU sampling variation. The binary outcome variable Y was generated from a logistic mixed-effects model including a random PSU effect to induce intra-class correlation (ICC):

$$\text{logit} \left[\Pr(Y_{hij} = 1) \right] = \beta_0 + \beta_1 X_{1,hij} + u_{hj}, \tag{9}$$

where $X_{1,hij} \sim \text{Bernoulli}(0.5)$ and $u_{hj} \sim \mathcal{N}(0, \sigma_u^2)$. The true parameter values were set to $\beta_0 = -0.85$ and $\beta_1 = \log(1.50)$, yielding a baseline prevalence of approximately 30%. The PSU random-effect variance σ_u^2 was derived from the ICC relationship:

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3}. \tag{10}$$

Three ICC levels (0.01, 0.05, and 0.10) represented weak, moderate, and strong clustering. Final analysis weights were constructed as the inverse of the stage-2 selection probabilities multiplied by an independent log-normal factor with variance parameter τ^2 , chosen to achieve a specified coefficient of variation of weights (CV_w). Weight-variability scenarios were defined by $CV_w = \{0.2, 0.5, 1.0\}$, representing low, moderate, and high dispersion. Crossing the three ICC levels with the three CV_w levels yielded nine design scenarios. Table 1 provides a concise overview of these nine combinations of ICC and weight-variability conditions.

Table 1. Summary of the nine simulation scenarios based on PSU size, ICC levels, and coefficient magnitudes.

Scenario	PSU Size	ICC Level	Coefficient Magnitude
1	Small	Low	Weak
2	Small	Medium	Moderate
3	Small	High	Strong
4	Medium	Low	Weak
5	Medium	Medium	Moderate
6	Medium	High	Strong
7	Large	Low	Weak
8	Large	Medium	Moderate
9	Large	High	Strong

For each scenario, 300 Monte Carlo replicates were generated to ensure stable estimates. Within each replicate, the stratified and clustered design was declared. All analyses were performed in R (version 4.4.2) using the survey package [13], following the framework outlined in Lumley’s comprehensive guide to complex survey analysis [18, 19]. The logistic regression of Y on X_1 was fitted using three variance-estimation approaches: TL via `svyglm(..., family = quasibinomial)`, Fay’s BRR with a shrinkage factor $\rho = 0.5$, and the Rao–Wu–Yue (RWY) rescaled bootstrap with $R = 80$ replicates. The number of bootstrap replicates ($R = 80$) was chosen to ensure stable variance estimates while maintaining reasonable computational cost. Prior evidence [5, 11] indicates that using between 50 and 100 replicates generally provides sufficiently accurate design-based variance estimates. Consistent with these recommendations, preliminary sensitivity checks in our study showed that increasing R beyond 80 changed the estimated standard errors by less than 1%, indicating that $R = 80$ offered an appropriate balance between precision and computational efficiency. For each fitted model, the point estimate, standard error (SE), confidence interval (CI) width, and coverage were recorded. A random seed of 20251011 was used to ensure reproducibility.

3.2 Performance Metrics

Each estimator was evaluated using four metrics: (i) 95% coverage, (ii) SE calibration, and (iii) median 95% CI width. For each replicate and each variance estimator, we recorded the point estimate of the target coefficient $\hat{\beta}_1$, its estimated SE \widehat{SE} , the two-sided 95% CI width $2z_{0.975}\widehat{SE}$, and a binary coverage indicator. Coverage at the 95% level was defined as the proportion of replicates for which

$$\hat{\beta}_1 \mp z_{0.975}\widehat{SE} \quad (11)$$

contained the true value $\beta_1 = \log(1.5)$, with $z_{0.975} = 1.96$. SE calibration was assessed by the ratio

$$\frac{\widehat{SE}}{SD_{MC}(\hat{\beta}_1)}, \quad (12)$$

where values near 1 indicate well-calibrated uncertainty. CI width was summarized as the median of the per-replicate 95% CI widths $2z_{0.975}\widehat{SE}$, a robust measure against outliers.

4 Simulation Results

4.1 Coverage Performance

Across all nine simulation scenarios combining different levels of clustering (ICC) and weight variability (CV_w), the three design-consistent variance estimators achieved coverage rates close to the nominal 95% level when clustering was weak or moderate ($ICC \leq 0.05$) (Table 2). These findings confirm the robustness of design-consistent inference under MICS-like two-stage survey structures. On average, coverage was 93.4% for TL, 93.1% for BRR, and 92.9% for the bootstrap. Coverage decreased slightly as the degree of clustering increased: averaging 95.0% for $ICC = 0.01$, 94.2% for $ICC = 0.05$, and 90.1% for $ICC = 0.10$, indicating increasing difficulty in variance approximation under stronger ICC. In contrast, weight heterogeneity exerted minimal influence on coverage, which remained near 93% across $CV_w = \{0.2, 0.5, 1.0\}$. At low clustering ($ICC = 0.01$), all estimators performed within the 94–96% range; under moderate clustering ($ICC = 0.05$), coverage stayed within 93–95%; and under strong clustering ($ICC = 0.10$), mild under-coverage of about 90% was observed. Overall, clustering, rather than weight variability, was the principal factor affecting coverage, and none of the three estimators consistently outperformed the others.

4.2 SE Calibration

For all ICC and CV_w combinations (Table 2), the ratio of the mean estimated SE to the Monte Carlo standard deviation of $\hat{\beta}_1$ (SE-ratio) remained close to unity across all estimators in most cases, confirming well-calibrated uncertainty quantification. Under weak clustering ($ICC = 0.01$), SE-ratios ranged from 0.98 to 1.03 across CV_w levels, indicating nominal accuracy.

At moderate clustering ($ICC = 0.05$), SE-ratios ranged between 0.93 and 1.01, while at strong clustering ($ICC = 0.10$) the ratios declined slightly to about 0.91–1.00, consistent with the minor under-coverage observed earlier. Averaged over all scenarios, SE-ratios were approximately 0.97 for TL, 0.97 for BRR, and 0.98 for the bootstrap, with differences negligible in practice. These results demonstrate that all three estimators yield well-calibrated SEs across a broad spectrum of clustering and weight heterogeneity, supporting their design-consistent validity and aligning with earlier simulation findings in the complex-survey literature [4, 5, 11].

4.3 CI Width Patterns

The 95% CI widths shown in Table 2 exhibit nearly identical results across estimators, confirming consistent precision and stable variance estimation. Variation in ICC produced minimal change in CI width, whereas increasing weight variability (CV_w) had a pronounced effect. The median CI widths were approximately 0.20 for $CV_w = 0.2$, 0.22 for $CV_w = 0.5$, and 0.27–0.28 for $CV_w = 1.0$, showing that weight heterogeneity contributes more to uncertainty than clustering. Three clear patterns emerged: (i) Increasing ICC from 0.01 to 0.10 had negligible impact on interval width, (ii) Increasing CV_w from 0.2 to 1.0 substantially widened CIs due to weight dispersion, and (iii) Differences among TL, BRR, and bootstrap were minimal, indicating that all estimators adequately capture the design effect when weights and clustering are correctly specified.

Overall, these results demonstrate that all estimators yield nearly identical precision under MICS-like two-stage survey structures, and that accounting for weight variability is far more influential than the specific choice of design-consistent variance estimators for achieving reliable inference.

5 Empirical Application: MICS Case Study

To complement the simulation analysis, an empirical application was conducted using data from the 2018–2019 UNICEF Multiple Indicator Cluster Survey (MICS) in the Central African Republic (CAR). The analysis focused on children aged 0–59 months with complete information on diarrhea status. Further survey details are available in the official report [20]. The MICS employed a two-stage stratified sampling design with regional stratification and clustering at the enumeration-area (PSU) level. Sampling weights accounted for unequal selection probabilities and nonresponse adjustments. The binary outcome variable, occurrence of diarrhea during the reference period, was modeled using a survey-weighted logistic regression framework. The explanatory variables included child sex, current breastfeeding status, weight-for-age z -score, and age group (0–11 [reference], 12–23, 24–35, 36–47, 48–59 months); maternal education (none/preschool [reference], primary, secondary or higher); type of residence (rural [reference], urban); region (1–7); household wealth

Table 2. Coverage (95%), SE calibration, and 95% CI width (R=300) by scenario and estimator

Scenario	Details (ICC, CV_w)	Estimator	Coverage (95%)	SE calib.	Lower CI	Upper CI
1	ICC=0.01, CV_w =0.20	Taylor	94.3%	0.98	91.7%	96.9%
		BRR	94.3%	0.98	91.7%	96.9%
		Bootstrap	95.0%	0.98	92.5%	97.5%
2	ICC=0.05, CV_w =0.20	Taylor	94.0%	0.94	91.3%	96.7%
		BRR	94.3%	0.93	91.7%	96.9%
		Bootstrap	93.0%	0.94	90.1%	95.9%
3	ICC=0.10, CV_w =0.20	Taylor	90.0%	0.99	86.6%	93.4%
		BRR	89.7%	0.99	86.3%	93.1%
		Bootstrap	90.0%	1.00	86.6%	93.4%
4	ICC=0.01, CV_w =0.50	Taylor	95.7%	1.03	93.4%	98.0%
		BRR	95.3%	1.03	92.9%	97.7%
		Bootstrap	95.3%	1.03	92.9%	97.7%
5	ICC=0.05, CV_w =0.50	Taylor	94.0%	1.00	91.3%	96.7%
		BRR	93.3%	1.00	90.4%	96.2%
		Bootstrap	93.7%	1.01	90.9%	96.5%
6	ICC=0.10, CV_w =0.50	Taylor	91.0%	0.97	87.8%	94.2%
		BRR	90.0%	0.97	86.6%	93.4%
		Bootstrap	89.7%	0.97	86.3%	93.1%
7	ICC=0.01, CV_w =1.00	Taylor	95.3%	1.00	92.9%	97.7%
		BRR	95.7%	1.00	93.4%	98.0%
		Bootstrap	94.3%	1.00	91.7%	96.9%
8	ICC=0.05, CV_w =1.00	Taylor	95.7%	0.98	93.4%	98.0%
		BRR	95.3%	0.98	92.9%	97.7%
		Bootstrap	94.7%	0.97	92.1%	97.2%
9	ICC=0.10, CV_w =1.00	Taylor	90.3%	0.91	87.0%	93.7%
		BRR	90.0%	0.91	86.6%	93.4%
		Bootstrap	90.0%	0.91	86.6%	93.4%

quintile (poorest [reference] to richest); source of drinking water (improved [reference] vs. unimproved); sanitation facility (improved [reference] vs. unimproved); and hand-washing facility (observed [reference] vs. unobserved). The survey-weighted logistic regression model was fitted under three design-consistent variance estimators: Taylor linearization (TL), Fay’s BRR, and the Rao–Wu–Yue (RWY) bootstrap (with $R = 80$ replicates). Estimation was performed in R using the survey package. For each method, odds ratios (ORs), 95% confidence intervals (CIs), and p -values were obtained. The results are summarized in Table 3.

As expected, the coefficient estimates (ORs) were identical across estimators, since point estimates in survey-weighted logistic regression are invariant to the chosen variance-estimation method. Differences arose only in the estimated SEs and CI widths. Replication-based estimators (BRR and bootstrap) produced slightly larger SEs and correspondingly wider, more conservative CIs—consistent with theoretical expectations [1, 4, 5]. Fay’s BRR approach stabilized variance estimates in strata with few PSUs, while the bootstrap effectively reproduced design-induced variability through re-sampling. Overall, these findings confirm that replication methods provide a more robust representation of design uncertainty under complex survey conditions, particularly when cluster effects or unequal weights are substantial.

6 Discussion and Conclusions

The present study confirms and extends classic findings in the literature on variance estimation for complex surveys. Early foundational works [3, 8] emphasized that ignoring clus-

tering and unequal weighting leads to underestimation of sampling variability, whereas replication-based approaches such as BRR and bootstrap [4, 5, 6] offer more stable inference under realistic survey conditions. Our results align with these principles but go further by providing a systematic simulation across a factorial grid of ICC and CV_w under MICS-like two-stage stratified designs, a combination rarely examined in prior studies.

Whereas previous comparisons typically evaluated each estimator in isolation or under simplified designs, this study quantifies the joint influence of clustering and weight heterogeneity on estimator performance. The results show that TL remains highly efficient under low-to-moderate design effects, but replicate-based methods, especially Fay’s BRR, retain nominal coverage when ICC and CV_w are high. This interaction effect between clustering and weight dispersion is a new contribution, emphasizing that weight variability, rather than clustering alone, drives the inflation of CIs and mild under-coverage. In this way, our simulation extends the studies of [9, 11] by empirically demonstrating when replication methods gain an advantage over linearization in multistage, unequal-weighted settings representative of UNICEF’s MICS, DHS, and LSMS surveys.

From a practical standpoint, the findings have direct implications for large-scale household surveys supporting the Sustainable Development Goals (SDGs), especially in low- and middle-income countries. Surveys like MICS employ multistage cluster designs with variable weights arising from regional stratification and nonresponse adjustments. In such contexts, TL remains efficient but may underestimate uncertainty for nonlinear or hierarchical indicators. Replication-based methods better capture design-induced variability, yielding slightly

Table 3. MICS survey-weighted logistic regression among the three design-consistent variance estimators.

Variable	TL					BRR					Bootstrap				
	ORs	L-CI	U-CI	SE	p-value	ORs	L-CI	U-CI	SE	p-value	ORs	L-CI	U-CI	SE	p-value
Sex: male	1.09	0.97	1.22	0.06	0.14	1.09	0.97	1.22	0.06	0.15	1.09	0.97	1.22	0.06	0.15
Breastfeeding: yes	1.31	1.11	1.55	0.09	0.0	1.31	1.09	1.57	0.09	0.0	1.31	1.11	1.54	0.09	0.0
Age: 12–23 months	2.29	1.77	2.95	0.13	0.0	2.29	1.76	2.97	0.13	0.0	2.29	1.76	2.96	0.13	0.0
Age: 24–35 months	1.78	1.41	2.24	0.12	0.0	1.78	1.37	2.30	0.13	0.0	1.78	1.39	2.29	0.12	0.0
Age: 36–47 months	1.27	0.98	1.65	0.13	0.07	1.27	0.98	1.65	0.13	0.07	1.27	0.98	1.65	0.13	0.07
Age: 48–59 months	1.22	0.94	1.57	0.13	0.13	1.22	0.92	1.60	0.14	0.16	1.22	0.92	1.59	0.14	0.14
Maternal education: primary	1.35	1.17	1.56	0.07	0.0	1.35	1.18	1.54	0.07	0.0	1.35	1.17	1.53	0.07	0.0
Maternal education: secondary or higher	1.39	1.13	1.71	0.10	0.0	1.39	1.15	1.69	0.10	0.0	1.35	1.14	1.70	0.10	0.0
Area: urban	1.03	0.80	1.33	0.13	0.82	1.03	0.79	1.34	0.13	0.83	1.03	0.79	1.33	0.13	0.82
Region: 2	0.97	0.75	1.25	0.13	0.82	0.97	0.78	1.21	0.11	0.79	0.97	0.80	1.20	0.11	0.78
Region: 3	1.19	0.90	1.59	0.15	0.22	1.19	0.90	1.58	0.14	0.21	1.19	0.90	1.58	0.14	0.21
Region: 4	1.28	1.01	1.62	0.12	0.04	1.28	0.99	1.65	0.13	0.06	1.28	1.10	1.63	0.13	0.06
Region: 5	1.08	0.81	1.45	0.15	0.59	1.08	0.83	1.41	0.13	0.55	1.08	0.82	1.42	0.14	0.56
Region: 6	1.53	1.20	1.94	0.12	0.0	1.53	1.18	1.98	0.13	0.0	1.53	1.19	1.97	0.13	0.0
Region: 7	0.97	0.70	1.35	0.17	0.87	0.97	0.69	1.36	0.17	0.87	0.97	0.69	1.36	0.17	0.87
Wealth index: poor	0.94	0.76	1.17	0.11	0.59	0.94	0.77	1.15	0.10	0.56	0.94	0.78	1.16	0.10	0.54
Wealth index: middle	0.99	0.80	1.23	0.11	0.94	0.99	0.83	1.18	0.09	0.92	0.99	0.82	1.22	0.09	0.93
Wealth index: rich	0.85	0.68	1.07	0.12	0.17	0.85	0.70	1.04	0.10	0.12	0.85	0.70	1.03	0.10	0.12
Wealth index: richest	0.90	0.64	1.26	0.17	0.54	0.90	0.63	1.28	0.18	0.55	0.90	0.64	1.28	0.17	0.54
Source of drinking water: unimproved	1.17	0.97	1.41	0.10	0.10	1.17	0.96	1.42	0.10	0.12	1.17	0.97	1.42	0.10	0.11
Sanitation status: unimproved	1.02	0.83	1.26	0.11	0.84	1.02	0.83	1.26	0.11	0.84	1.02	0.84	1.26	0.11	0.84
Handwashing facilities: unobserved	0.82	0.69	0.96	0.08	0.02	0.82	0.69	0.96	0.08	0.02	0.82	0.69	0.96	0.08	0.02

wider yet more credible CIs.

For analysts using MICS data, the study recommends employing TL for descriptive indicators, Fay’s BRR when PSUs per stratum are few or unbalanced (with $\rho \approx 0.5$), and the bootstrap for nonlinear or small area models. These recommendations support global data–harmonization initiatives led by UNICEF, WHO, and the World Bank, emphasizing that the choice of design–consistent variance estimator should reflect the true design complexity rather than convenience, thereby ensuring methodological rigor in policy–relevant survey analyses.

Our results translate into several scenario-specific recommendations. When ICC is high and CV_w is low, strong within-PSU homogeneity implies that clustering dominates the design effect; thus, design-based or multilevel models are essential to avoid underestimated variances. When ICC is low but CV_w is high, unequal weighting becomes the primary driver of estimator instability, making correct weight incorporation and potential weight stabilization crucial. Small or highly unbalanced strata further complicate variance estimation, where replication-based methods generally outperform TL. For nonlinear estimators, including logistic regression with interactions and modern machine learning models, the curvature of the estimation problem magnifies sensitivity to weight extremes and cluster size disparities. Accordingly, analysts should employ robust variance estimators, conduct sensitivity analyses, and interpret performance differences with caution.

Limitations of the Simulation Assumptions

Our simulation framework was intentionally designed to reflect key elements of MICS/DHS survey structures; however,

several simplifying assumptions may limit the generalizability of the findings. First, we assumed equal PSU sizes and balanced strata, whereas real-world household surveys often exhibit substantial variation in cluster population, sampling fractions, and nonresponse patterns. Such heterogeneity can influence both estimator stability and variance behavior. Second, the data-generating mechanism relied on a single logistic regression model. Although logistic models are widely used in analyses of MICS/DHS datasets, alternative outcome structures (e.g., multinomial or count models, hierarchical processes, or non-linear machine-learning mechanisms) may behave differently under complex designs. These assumptions were adopted to maintain computational tractability and isolate the performance of the estimators under controlled conditions, but they also narrow the scope of extrapolation to highly irregular or heterogeneous survey contexts. Future work should extend the design to incorporate unequal cluster sizes, unbalanced strata, and multiple data-generating mechanisms to more fully represent the diversity of survey environments encountered in practice.

Data Availability

This study used de-identified secondary data from the 2018–2019 UNICEF Multiple Indicator Cluster Survey (MICS), which are publicly accessible at <https://mics.unicef.org/>. Ethical approval and informed consent were obtained during the original data collection by the implementing agencies. The secondary analysis of anonymized data involved no human interaction and posed no risk to participant privacy or confidentiality.

Conflicts of Interest

The authors declare no conflicts of interest.

10.1177/1536867X1001000201.

REFERENCES

- [1] Heeringa S. G., West B. T., Berglund P. A., “Applied survey data analysis,” in *Applied Survey Data Analysis*, 2nd ed., Chapman & Hall/CRC, Boca Raton, 2017.
- [2] Lohr S. L., “Sampling: design and analysis,” in *Sampling: Design and Analysis*, 3rd ed., CRC Press, Boca Raton, 2022.
- [3] Wolter K. M., “Introduction to variance estimation,” in *Introduction to Variance Estimation*, 2nd ed., Springer, New York, 2007.
- [4] Rust K. F., Rao J. N. K., “Variance estimation for complex surveys using replication techniques,” *Statistical Methods in Medical Research*, vol. 5, no. 3, pp. 283–310, 1996. DOI: 10.1177/096228029600500305.
- [5] Rao J. N. K., Wu C. F. J., “Resampling inference with complex survey data,” *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 231–241, 1988. DOI: 10.1080/01621459.1988.10478591.
- [6] Sitter R. R., “A resampling procedure for complex survey data,” *Journal of the American Statistical Association*, vol. 87, no. 419, pp. 755–765, 1992. DOI: 10.1080/01621459.1992.10476263.
- [7] Särndal C.-E., Swensson B., Wretman J., “Model assisted survey sampling,” in *Model Assisted Survey Sampling*, Springer, New York, 1992.
- [8] Kish L., Frankel M. R., “Inference from complex samples,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 1, pp. 1–37, 1974. <https://www.jstor.org/stable/2984767>
- [9] Beaumont J.-F., Charest A.-S., “Bootstrap variance estimation with survey data when estimating model parameters,” *Computational Statistics and Data Analysis*, vol. 56, no. 12, pp. 4450–4461, 2012. DOI: 10.1016/j.csda.2012.04.015.
- [10] Fay M. P., Graubard B. I., “Small-sample adjustments for Wald-type tests using sandwich estimators,” *Biometrics*, vol. 57, no. 4, pp. 1198–1206, 2001. DOI: 10.1111/j.0006-341X.2001.01198.x.
- [11] Kolenikov S., “Resampling variance estimation for complex survey data,” *The Stata Journal: Promoting Communications on Statistics and Stata*, vol. 10, no. 2, pp. 165–199, 2010. DOI: 10.1177/1536867X1001000201.
- [12] Binder D. A., “On the variances of asymptotically normal estimators from complex surveys,” *International Statistical Review*, vol. 51, no. 3, pp. 279–292, 1983. DOI: 10.2307/1402588.
- [13] Lumley T., “The survey package for R: analysis of complex survey samples,” *CRAN Package Documentation*. <https://cran.r-project.org/web/packages/survey/index.html> (retrieved Feb. 7, 2025).
- [14] Chatterjee S., Bose A., “Generalized bootstrap for estimating equations,” *Annals of Statistics*, vol. 33, no. 1, pp. 414–436, 2005. DOI: 10.1214/009053604000001505.
- [15] Davison A. C., Hinkley D. V., “Bootstrap methods and their application,” in *Bootstrap Methods and Their Application*, Cambridge Univ. Press, 1997.
- [16] Shao J., Tu D., “The jackknife and bootstrap,” in *The Jackknife and Bootstrap*, Springer, New York, 1995.
- [17] Beaumont J.-F., “A bootstrap variance estimation method for multistage sampling designs,” *Stats*, vol. 5, no. 2, pp. 228–256, 2022. DOI: 10.3390/stats5020017.
- [18] Lumley T., “Complex surveys: a guide to analysis using R,” in *Complex Surveys: A Guide to Analysis Using R*, Wiley, Hoboken, NJ, 2010.
- [19] Lumley T., “Analysis of complex survey samples,” *Journal of Statistical Software*, vol. 9, no. 8, pp. 1–19, 2004. DOI: 10.18637/jss.v009.i08.
- [20] ICASEES/UNICEF, “MICS6–RCA Multiple Indicator Cluster Survey 2018–2019: Final Report of Survey Results,” UNICEF MICS Website. <https://mics.unicef.org/surveys> (retrieved Feb. 7, 2025).