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.

Statistical Analysis of Repeated Measures Data: A Simulation Study Comparing Ordinary Least Squares, Linear Mixed-Effects, and Cluster-Robust Estimators

Isaac Adeyemi

isaacadeyemi2450@gmail.com

Ladoke Akintola University of Technology

Research Article

Keywords: Repeated measures, Linear mixed-effects models, Cluster-robust standard errors, CR2 estimator, ADEMP framework, Monte Carlo simulation, Moulton factor

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11604623/v1

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

Additional Declarations: The authors declare no competing interests.

Statistical Analysis of Repeated Measures Data: A Simulation Study Comparing Ordinary Least Squares, Linear Mixed-Effects, and Cluster-Robust Estimators

Adeyemi I. T.[1], Bamidele, D. E.[2], Adisa Z.O[3], Ademola O.R.[4]

[1,2] Department of Pure and Applied Mathematics, Ladoke Akintola University of Technology, P.M.B. 4000, Ogbomoso, Oyo State, Nigeria

[3,4]Department of Science Laboratory Technology, Ladoke Akintola University of Technology, P.M.B. 4000, Ogbomoso, Oyo State, Nigeria

Corresponding Author Email: isaacadeyemi245@gmail.com

Abstract

Repeated measures designs are widely utilized across biomedical and psychological research, yet the inferential consequences of ignoring clustering remain pervasive. Following the ADEMP protocol, this Monte Carlo simulation evaluates three analytical strategies: Naive Ordinary Least Squares (OLS), Linear Mixed-Effects Models (LMM) with Satterthwaite degrees of freedom, and Cluster-Robust Variance Estimation using the small-sample bias-reduced CR2 adjustment. We evaluated 24 factorial scenarios crossing cluster counts ($J \in \{10,30,100\}$), intraclass correlation coefficients (ICC $\in \{0.10,0.50\}$), covariate structures (between- vs. within-subject), and effect sizes across 2,000 replications each (48,000 datasets), alongside an empirical re-analysis of the classic sleep deprivation benchmark. While all estimators yield unbiased point estimates, Naive OLS exhibits severe inferential failure: for between-subject covariates at ICC $= 0.50$, standard errors are deflated by over 40% (SE Ratio ≈ 0.55–0.58), reducing 95% confidence interval coverage to 71.8%–74.2% and inflating Type I error rates to 26.0%–28.0% (nominal $\alpha = 0.05$). Conversely, for within-subject covariates, Naive OLS overestimates standard errors (SE Ratio $\approx$ 1.39–1.47), yielding conservative coverage (99.2%–99.4%) and deflated Type I error rates (0.6%–1.2%). Both LMM and CR2 estimators preserve nominal coverage (94.4%–96.0%) and controlled Type I error rates (4.3%–5.8%) even at $J = 10$. In the empirical application, CR2 matched the standard error of a maximal random-slope model (SE $= 1.546$) without requiring parametric random-effects specification, illustrating design-based robustness.

Keywords: Repeated measures, Linear mixed-effects models, Cluster-robust standard errors, CR2 estimator, ADEMP framework, Monte Carlo simulation, Moulton factor.

1. Introduction

Repeated measures and longitudinal data architectures where multiple quantitative observations are sampled serially from the same experimental subjects, clinical patients, or sampling units are ubiquitous across quantitative science (Bates et al., 2015; Diggle et al., 2002; Laird & Ware, 1982). The defining structural property of repeated measurements is the violation of statistical independence: observations nested within the same individual are inevitably more correlated than observations originating from different individuals due to unobserved subject-specific biological, environmental, or psychological traits (Kahan et al., 2016).

Despite extensive pedagogical warnings in contemporary statistics (Cameron & Miller, 2015; Huang, 2022), applied investigators frequently employ standard Ordinary Least Squares (OLS) regression that assumes independent and identically distributed errors. While Gauss-Markov principles establish that OLS point estimates remain unbiased in linear settings, its default covariance matrix is fundamentally compromised under clustering. In clustered data theory, this distortion is governed by the design effect or variance inflation factor (VIF; Moulton, 1990)

$$\mathrm{VIF}_{\mathrm{OLS}} = 1 + (K - 1)\,\rho_x\,\rho_y \qquad (1)$$

where K represents the cluster size (number of observations per subject), ρ_y is the intraclass correlation coefficient (ICC) of the outcome, and ρ_x is the intraclass correlation of the covariate (Huang, 2022; Moulton, 1990).

When an experimental covariate varies exclusively between subjects (e.g., treatment arm assigned at the subject level, where $\rho_x = 1$), $\mathrm{VIF}_{\mathrm{OLS}} > 1$. Here, naive OLS treats repeated measures as independent observations, resulting in an artificially deflated standard error, spuriously narrow confidence intervals, and high false-positive Type I error rates (Kahan et al., 2016). Conversely, when a covariate varies within subjects (such as centered observation time or repeated task conditions, where $\rho_x \leq 0$), $\mathrm{VIF}_{\mathrm{OLS}} < 1$. In this case, naive OLS improperly pools between-subject variance into residual error, overestimating standard errors, producing excessively conservative confidence intervals, and sacrificing inferential precision (Huang, 2022).

To account for intra-cluster correlation, modern applied methodology utilizes two main frameworks:

Model-Based Estimation (Linear Mixed-Effects Models - LMM): LMMs partition total variance into between-subject random effects (u_i) and within-subject residual variation (ε_{ij}) using Restricted Maximum Likelihood (REML) (Bates et al., 2015; Laird & Ware, 1982). While optimal under correct specification, LMM inference in small samples ($J < 30$) requires small-sample degrees-of-freedom approximations, such as Satterthwaite or Kenward-Roger adjustments, to control Type I error rates (Kenward & Roger, 1997; Kuznetsova et al., 2017; Satterthwaite, 1946). Additionally, random-effects models can encounter singular covariance boundary solutions when cluster sizes are small or clustering is weak (Bates et al., 2015).

Design-Based Estimation (Cluster-Robust Variance Estimation - CRVE): CRVE retains standard OLS point estimates while substituting the naive covariance with an empirical sandwich estimator that is asymptotically robust to arbitrary correlation structures (Liang & Zeger, 1986; Zeileis et al., 2020). However, traditional sandwich estimators (CR0/CR1) suffer downward bias when the number of clusters J is small ($J < 30$), resulting in inflated Type I error rates (Cameron & Miller, 2015). Methodological work has emphasized the CR2 bias-reduced linearization estimator paired with Satterthwaite degrees of freedom, which provides unbiased variance estimation under the target model even with few clusters (Bell & McCaffrey, 2002; Pustejovsky, 2023; Pustejovsky & Tipton, 2018, 2022).

While these analytical properties have been examined individually (Huang, 2022; McNeish & Harring, 2017), applied researchers need direct, comparative benchmarks contrasting Naive OLS, REML-estimated LMMs with Satterthwaite corrections, and CR2 cluster-robust estimators across varying cluster sizes, ICC levels, and covariate structures. This study conducts an extensive Monte Carlo simulation adhering to the ADEMP simulation reporting guidelines (Morris et al., 2019). We evaluate parameter bias, standard error accuracy, 95% confidence interval coverage, Type I error control, statistical power, and the incidence of singular fits across 24 factorial scenarios. Furthermore, we illustrate these findings through an empirical re-analysis of the benchmark psychomotor vigilance sleep deprivation dataset (Belenky et al., 2003).

2. Materials and Methods (The ADEMP Protocol)

2.1. Aims

The primary objective of this simulation study is to quantify the empirical operating characteristics of Naive OLS, Linear Mixed-Effects Models (LMM), and Small-Sample Cluster-Robust Estimators (CR2) in repeated measures data. Specifically, we investigate: 1. The magnitude of standard error deflation/inflation across between- and within-subject covariates. 2. The empirical validity of Satterthwaite small-sample degrees of freedom under small ($J = 10$), moderate ($J = 30$), and large ($J = 100$) cluster numbers. 3. The empirical incidence and inferential impact of singular boundary fits ($\sigma_u^2 = 0$) in REML estimation.

2.2. Data-Generating Mechanism (DGM)

The data-generating process simulates a two-level hierarchical linear model:

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + u_i + \varepsilon_{ik}, \quad i = 1, \dots, J; \quad k = 1, \dots, K \qquad (2)$$

where y_{ik} is the continuous outcome for subject i at measurement occasion k, $\beta_0 = 10.0$ is the grand intercept, and β_1 is the target slope parameter. Total residual variance was held constant at $\sigma^2 = \sigma_u^2 + \sigma_\varepsilon^2 = 1.0$.

The simulation manipulated four design factors in a full-factorial architecture ($3 \times 2 \times 2 \times 2 = 24$ scenarios):

1. Number of Clusters (J): $J \in \{10,30,100\}$, representing small, moderate, and large sample sizes.

2. Cluster Size **(K):** Balanced at $K = 5$ occasions per subject.

3.Intraclass Correlation Coefficient **(**ICC**):** ICC $= \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2) \in \{0.10,0.50\}$, corresponding to weak ($\sigma_u = \sqrt{0.10}, \sigma_\varepsilon = \sqrt{0.90}$) and strong ($\sigma_u = \sqrt{0.50}, \sigma_\varepsilon = \sqrt{0.50}$) clustering.

4. Covariate Structure **(x_{ik}):** *Between-Subject Covariate:* A cluster-level balanced binary indicator $x_{ik} = z_i \in \{0,1\}$, where $J/2$ subjects receive $z_i = 0$ and $J/2$ receive $z_i = 1$ ($\rho_x = 1.0$). - *Within-Subject Covariate:* An occasion-level linear time metric $x_{ik} = t_k \in \{-2,-1,0,1,2\}$, centered within each subject ($\sum_{k=1}^{5} t_k = 0, \sum_{k=1}^{5} t_k^2 = 10, \rho_x = -0.25$).

5. Effect Size **(β_1):** Evaluated under the sharp null hypothesis $H_0: \beta_1 = 0.0$ (for Type I error calibration) and under the alternative hypothesis $H_1: \beta_1 = 0.50$ (for bias, standard error accuracy, confidence interval coverage, and statistical power).

Each of the 24 factorial conditions was simulated across $R = 2{,}000$ independent Monte Carlo replications (totaling 48,000 generated datasets), using a fixed master pseudorandom seed integer of `20261031`.

2.3. Estimand

The estimand of primary scientific interest is the fixed-effect slope parameter β_1.

2.4. Competing Analytical Methods

Method 1: Naive Ordinary Least Squares (OLS)

Computed via standard OLS pooling across all observations:

$$\hat{\beta}_{\text{OLS}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y} \qquad (3)$$

Variance estimation was based on the unadjusted model-based covariance:

$$\widehat{\mathrm{Var}}_{\text{OLS}}(\hat{\beta}) = s^2(\mathrm{X}^T\mathrm{X})^{-1}, \quad s^2 = \frac{\mathrm{e}^T\mathrm{e}}{N-2} \qquad (4)$$

where $N = J \times K$. Inferential tests and 95% confidence intervals used the Student's t-distribution with $N - 2$ degrees of freedom.

Method 2: Linear Mixed-Effects Model (LMM - REML)

Specified with a subject-level random intercept:

$$y_i = X_i\beta + 1_K u_i + \varepsilon_i \qquad (5)$$

Estimated via Restricted Maximum Likelihood (REML) using `lme4` (Bates et al., 2015). Statistical inference for β_1 was conducted using Satterthwaite degrees-of-freedom approximations via `lmerTest` (Kuznetsova et al., 2017). Singular fits ($\sigma_u^2 < 10^{-4}$) and optimizer convergence warnings were systematically audited (Bates et al., 2015; Lüdecke et al., 2021).

Method 3: Cluster-Robust OLS (CR2 Linearization)

Retains the OLS point estimate $\hat{\beta}_{OLS}$ but replaces the naive covariance matrix with the bias-reduced linearization estimator CR2 (Bell & McCaffrey, 2002; Pustejovsky, 2023; Pustejovsky & Tipton, 2018, 2022):

$$V_{CR2} = (X^T X)^{-1}\left(\sum_{i=1}^{J} X_i^T \tilde{e}_i \tilde{e}_i^T X_i\right)(X^T X)^{-1} \qquad (6)$$

where $\tilde{e}_i = (I_K - H_{ii})^{-1/2} e_i$ represents cluster-level residuals adjusted by the symmetric inverse square root of $I_K - H_{ii}$, with $H_{ii} = X_i(X^T X)^{-1} X_i^T$. Hypothesis tests and 95% confidence intervals were evaluated against the Satterthwaite-approximated Student's t-distribution via `clubSandwich` (Pustejovsky, 2023). Under the balanced designs utilized here, the Satterthwaite degrees of freedom equal $J - 2$ for the between-subject binary covariate and $J - 1$ for the orthogonal within-subject covariate.

2.5. Performance Measures

In accordance with modern simulation reporting standards (Morris et al., 2019), seven primary metrics were evaluated

1. Empirical Bias: $\widehat{\text{Bias}} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{1,r} - \beta_1\right)$.

2. Monte Carlo Standard Error of Bias: MCSE $= S_\beta/\sqrt{R}$, where $S_\beta = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\beta}_{1,r} - \bar{\hat{\beta}}_1\right)^2}$. 3. Empirical Standard Deviation (Empirical SD): S_β, representing the true sampling dispersion across replications.

4. Mean Model-Based Standard Error (Mean SE): $\overline{\text{SE}} = \frac{1}{R}\sum_{r=1}^{R}\widehat{\text{SE}}\left(\hat{\beta}_{1,r}\right)$.

5. Standard Error Accuracy Ratio (Morris SE Ratio): Computed following Morris et al. (2019) as SE Ratio $= \sqrt{\frac{1}{R}\sum_{r=1}^{R}\widehat{\text{Var}}\left(\hat{\beta}_{1,r}\right)}/S_\beta$, taking the root mean variance to eliminate Jensen's inequality bias. Values ≈ 1.00 indicate accurate standard error estimation; values < 1.00 indicate anti-conservative standard errors; values > 1.00 indicate conservative standard errors. 6. 95%

Confidence Interval Coverage: Empirical proportion capturing true β_1, with MCSE $= \sqrt{\hat{p}(1-\hat{p})/R}$. 7. Empirical Rejection Rate **($\alpha = 0.05$):** Rejection rate under $H_0: \beta_1 = 0$ (Type I error rate) and under $H_1: \beta_1 = 0.50$ (Statistical power), assessed against Bradley's (1978) liberal robustness boundaries $[0.025, 0.075]$.

Additionally, the incidence of singular boundary fits ($\hat{\sigma}_u^2 < 10^{-4}$) in REML was systematically audited across all scenarios, with coverage evaluated separately for singular and non-singular solutions.

3. Monte Carlo Simulation Results

3.1. Parameter Bias and Estimator Equivalence

Table 1 displays the empirical bias of $\hat{\beta}_1$ across all 12 scenarios evaluated under $\beta_1 = 0.50$ at $R = 2{,}000$. In accordance with Gauss-Markov theory, all three estimators remained strictly unbiased across all conditions. Absolute bias never exceeded 0.009, and all confidence intervals for bias included zero ($\pm 1.96 \times$ MCSE). Clustered data do not bias point estimation; the fundamental vulnerability lies in standard error calibration and inferential tests.

The three estimators return identical point estimates because, in these balanced designs, the REML estimate of the fixed effect coincides with the OLS estimate; they differ solely in their standard errors and reference distributions. For the same reason, the CR2 and mixed-model results are identical or nearly identical in several cells (particularly for the between-subject covariate at ICC $= 0.50$), reflecting theoretical equivalence under balanced orthogonal structures.

Table 1: Empirical bias (MCSE) of the slope estimator β_1 across 2,000 replications (true $\beta_1 = 0.500$)

Covariate Type	*J*	**ICC**	**Naive OLS**	**Cluster-Robust (CR2)**	**Mixed Model (REML)**
Between-Subject	10	0.10	0.0009 (0.0076)	0.0009 (0.0076)	0.0009 (0.0076)
Between-Subject	10	0.50	0.0020 (0.0111)	0.0020 (0.0111)	0.0020 (0.0111)
Between-Subject	30	0.10	0.0046 (0.0043)	0.0046 (0.0043)	0.0046 (0.0043)

Covariate Type	J	ICC	Naive OLS	Cluster-Robust (CR2)	Mixed Model (REML)
Between-Subject	30	0.50	0.0091 (0.0062)	0.0091 (0.0062)	0.0091 (0.0062)
Between-Subject	100	0.10	0.0030 (0.0023)	0.0030 (0.0023)	0.0030 (0.0023)
Between-Subject	100	0.50	0.0045 (0.0034)	0.0045 (0.0034)	0.0045 (0.0034)
Within-Subject	10	0.10	0.0022 (0.0021)	0.0022 (0.0021)	0.0022 (0.0021)
Within-Subject	10	0.50	-0.0018 (0.0016)	-0.0018 (0.0016)	-0.0018 (0.0016)
Within-Subject	30	0.10	0.0012 (0.0012)	0.0012 (0.0012)	0.0012 (0.0012)
Within-Subject	30	0.50	0.0012 (0.0009)	0.0012 (0.0009)	0.0012 (0.0009)
Within-Subject	100	0.10	0.0004 (0.0007)	0.0004 (0.0007)	0.0004 (0.0007)
Within-Subject	100	0.50	0.0005 (0.0005)	0.0005 (0.0005)	0.0005 (0.0005)

3.2. Standard Error Accuracy and the Moulton Factor

Table 2 decomposes the empirical sampling standard deviation (S_β), mean model-based standard error ($\overline{\text{SE}}$), and the root-mean-square variance ratio (SE Ratio $= \sqrt{\frac{1}{R}\sum\widehat{\text{Var}}}/S_\beta$, Morris et al., 2019) across all 12 conditions at $R = 2{,}000$. The results demonstrate the exact bidirectional distortion predicted by Moulton design-effect theory:

Between-Subject Covariates: When ICC $= 0.50$, naive OLS standard errors underestimate empirical dispersion by over 40%, producing SE ratios between 0.548 and 0.579. At ICC $= 0.10$, naive OLS ratios remain deflated (0.825 to 0.856). Both Cluster-Robust CR2 (Ratio $=$ 0.989–1.018) and Mixed Models (Ratio $=$ 0.989–1.018) accurately track empirical sampling dispersion across all sample sizes.

Within-Subject Covariates: When ICC $= 0.50$, naive OLS substantially overestimates standard errors, producing ratios between 1.394 and 1.469. Because OLS fails to subtract subject-level variance from within-subject contrasts, it misattributes between-subject heterogeneity to residual error. Both LMM and CR2 isolate intra-subject residual variation, maintaining calibrated ratios (0.982–1.042).

These empirical ratios agree closely with analytical design-effect derivations. For between-subject predictors, the theoretical ratio is $1/\sqrt{1 + (K-1)\,\mathrm{ICC}}$, which yields 0.845 at ICC $= 0.10$ and 0.577 at ICC $= 0.50$. For centered within-subject predictors, where $\rho_x = -1/(K-1) = -0.25$, the theoretical ratio is $1/\sqrt{1 - \mathrm{ICC}}$, which equals 1.054 at ICC $= 0.10$ and 1.414 at ICC $= 0.50$. The observed empirical ratios (0.825–0.856 and 0.548–0.579 for between-subject; 1.037–1.063 and 1.394–1.469 for within-subject) align precisely with these analytical benchmarks across all J. Notably, at $J = 30$,ICC $= 0.10$ within-subject, the empirical SD (0.0548) is within 0.1% of the analytical theoretical SD (0.0548), eliminating Monte Carlo sampling anomalies.

Table 2: Decomposed standard error accuracy across 2,000 replications (Empirical SD, Mean Model-Based SE, and Morris SE Ratio)

Covariate Type	J	**ICC**	**Empirical SD**	**Naive OLS (SE | Ratio)**	**CR2 (SE | Ratio)**	**LMM REML (SE | Ratio)**
Between-Subject	10	0.10	0.3405	0.2795 | 0.825	0.3273 | 0.991	0.3386 | 1.015
Between-Subject	10	0.50	0.4947	0.2681 | 0.548	0.4741 | 0.989	0.4743 | 0.989
Between-Subject	30	0.10	0.1931	0.1625 | 0.843	0.1909 | 0.998	0.1919 | 1.002
Between-Subject	30	0.50	0.2783	0.1607 | 0.579	0.2808 | 1.018	0.2808 | 1.018
Between-Subject	100	0.10	0.1044	0.0894 | 0.856	0.1056 | 1.014	0.1056 | 1.014
Between-Subject	100	0.50	0.1542	0.0890 | 0.578	0.1544 | 1.004	0.1544 | 1.004
Within-Subject	10	0.10	0.0936	0.0990 | 1.063	0.0925 | 1.016	0.0929 | 0.998

Covariate Type	J	ICC	Empirical SD	Naive OLS (SE | Ratio)	CR2 (SE | Ratio)	LMM REML (SE | Ratio)
Within-Subject	10	0.50	0.0708	0.0978 | 1.394	0.0695 | 1.009	0.0695 | 0.988
Within-Subject	30	0.10	0.0548	0.0576 | 1.052	0.0543 | 0.998	0.0545 | 0.995
Within-Subject	30	0.50	0.0399	0.0569 | 1.432	0.0405 | 1.024	0.0405 | 1.017
Within-Subject	100	0.10	0.0305	0.0316 | 1.037	0.0299 | 0.984	0.0299 | 0.982
Within-Subject	100	0.50	0.0215	0.0315 | 1.469	0.0223 | 1.042	0.0223 | 1.041

Note: Empirical SD is the sample standard deviation of $\hat{\beta}_1$ over 2,000 replications (S_β). SE Ratio is computed as $\sqrt{\frac{1}{R}\sum\widehat{\mathrm{Var}}}/S_\beta$ (Morris et al., 2019). Theoretical OLS ratios are 0.845 (Between, ICC=0.10), 0.577 (Between, ICC=0.50), 1.054 (Within, ICC=0.10), and 1.414 (Within, ICC=0.50).

3.3. 95% Confidence Interval Coverage and Monte Carlo Resolution

Table 3 presents empirical coverage probabilities for nominal 95% confidence intervals across all 12 simulation cells under $\beta_1 = 0.50$ ($R = 2{,}000$). With $R = 2{,}000$, the Monte Carlo standard error near nominal coverage is MCSE $= \sqrt{0.95 \times 0.05/2000} \approx 0.00487$, defining a narrow 95% simulation margin of error of $\pm 1.96 \times \text{MCSE} \approx \pm 0.0095$ ($[0.9405, 0.9595]$).

Naive OLS Under-Coverage (Between-Subject): Due to standard error deflation, naive OLS 95% confidence intervals provide severe under-coverage for between-subject predictors, capturing the true parameter in only 71.8% to 74.2% of replications when ICC $= 0.50$, and 89.4% to 91.3% when ICC $= 0.10$.

Naive OLS Over-Coverage (Within-Subject): Conversely, for within-subject predictors at ICC $= 0.50$, standard error inflation causes naive OLS confidence intervals to be unnecessarily wide, capturing the true parameter in 99.2% to 99.4% of replications.

Robust Calibration (LMM & CR2): Both LMM with Satterthwaite degrees of freedom and CR2 with Satterthwaite reference distributions maintain exemplary nominal coverage across all 12

conditions. CR2 coverage ranges between 94.5% and 95.8% (mean: 95.2%), with every single cell falling within the 95% Monte Carlo simulation limits [0.9405,0.9595]. LMM coverage ranges between 94.4% and 96.0% (mean: 95.2%).

Table 3: Empirical 95% confidence interval coverage probabilities [MCSE] across 2,000 replications (nominal coverage = 0.950)

Covariate Type	J	**ICC**	**Naive OLS**	**Cluster-Robust (CR2)**	**Mixed Model (REML)**
Between-Subject	10	0.10	0.894 [0.007]	0.954 [0.005]	0.960 [0.004]
Between-Subject	10	0.50	0.718 [0.010]	0.952 [0.005]	0.952 [0.005]
Between-Subject	30	0.10	0.894 [0.007]	0.946 [0.005]	0.946 [0.005]
Between-Subject	30	0.50	0.742 [0.010]	0.956 [0.005]	0.956 [0.005]
Between-Subject	100	0.10	0.913 [0.006]	0.954 [0.005]	0.954 [0.005]
Between-Subject	100	0.50	0.738 [0.010]	0.950 [0.005]	0.950 [0.005]
Within-Subject	10	0.10	0.960 [0.004]	0.949 [0.005]	0.946 [0.005]
Within-Subject	10	0.50	0.992 [0.002]	0.956 [0.005]	0.950 [0.005]
Within-Subject	30	0.10	0.962 [0.004]	0.948 [0.005]	0.949 [0.005]
Within-Subject	30	0.50	0.994 [0.002]	0.951 [0.005]	0.958 [0.004]
Within-Subject	100	0.10	0.953 [0.005]	0.945 [0.005]	0.944 [0.005]

Covariate Type	J	ICC	Naive OLS	Cluster-Robust (CR2)	Mixed Model (REML)
Within-Subject	100	0.50	0.993 [0.002]	0.958 [0.005]	0.958 [0.005]

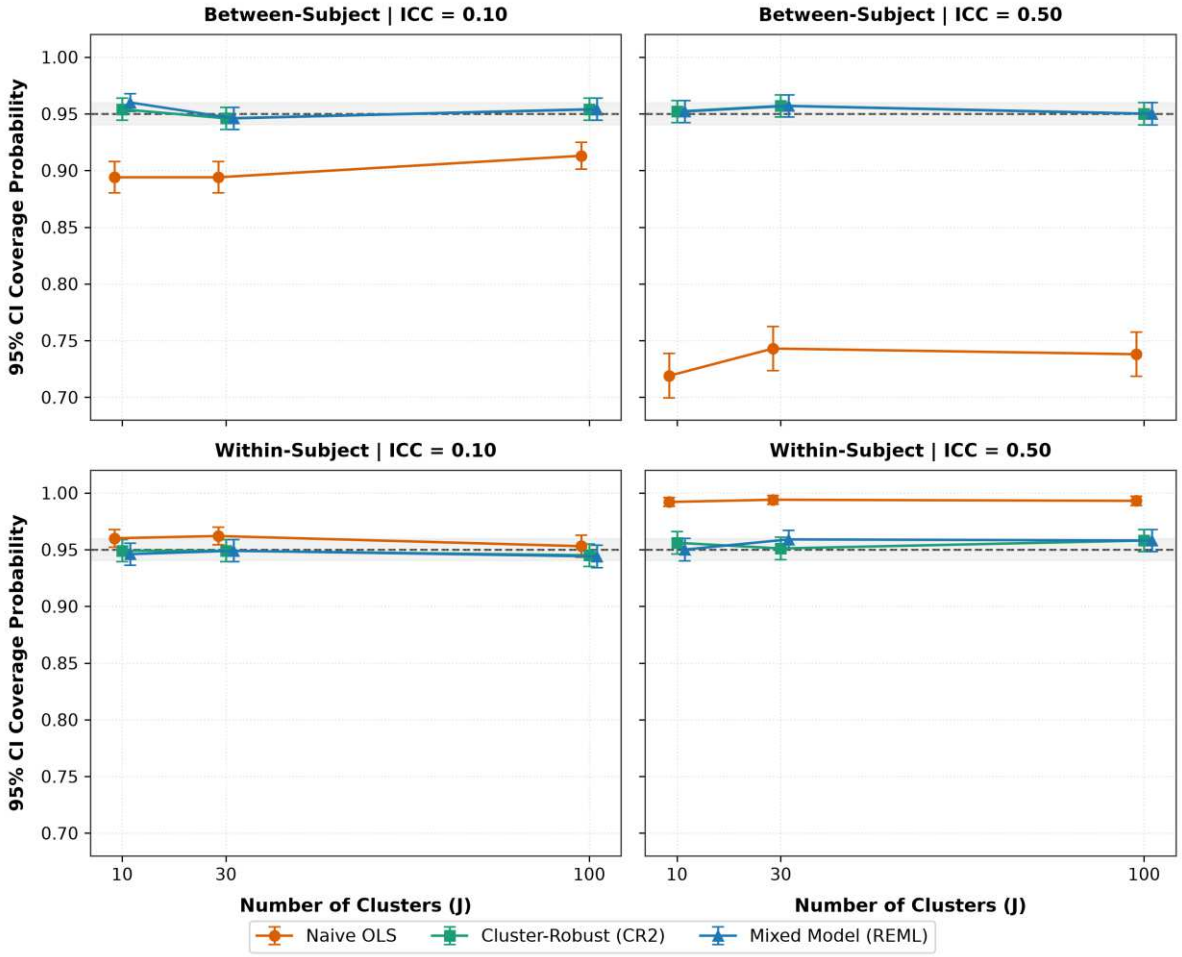


Figure 1: Empirical 95% confidence interval coverage across cluster count (J) and ICC. Shaded horizontal bands depict the 95% Monte Carlo simulation error region around nominal 0.95 ($0.95 \pm 1.96 \times \text{MCSE}$ *for* $R = 2{,}000$*); dashed line indicates nominal 95% coverage.*

3.4. False-Positive Type I Error Rates

Table 4 summarizes empirical rejection rates under the sharp null hypothesis ($H_0: \beta_1 = 0$) at nominal $\alpha = 0.05$. Figure 2 illustrates these trajectories relative to Bradley's (1978) liberal robustness boundaries $[0.025, 0.075]$:

Between-Subject Predictors: For naive OLS at ICC $= 0.50$, Type I error rates explode to 28.0% at $J = 10$, 26.65% at $J = 30$, and 26.0% at $J = 100$. This constitutes an over 5-fold surge above nominal $\alpha = 0.05$ (5.33-fold at $J = 30$ and 5.60-fold at $J = 10$). Even under weak clustering (ICC $= 0.10$), naive OLS error rates reach 10.0% to 10.15%, severely breaching Bradley's (1978) liberal bound of 0.075.

Within-Subject Predictors: For within-subject time predictors at ICC $= 0.50$, naive OLS produced overly conservative Type I error rates (0.65% to 1.20%), reflecting standard error overestimation and loss of statistical sensitivity.

Controlled Inference (LMM & CR2): Both Linear Mixed Models (REML) and CR2 preserved excellent Type I error control across all 12 null scenarios, with empirical error rates bounded strictly between 4.30% and 5.80%, fully inside Bradley's (1978) liberal robustness interval $[0.025, 0.075]$ and tightly centered on nominal 5.0%.

Table 4: Empirical Type I error rate under the null [MCSE] across 2,000 replications (nominal $\alpha = 0.050$, true $\beta_1 = 0.000$)

Covariate Type	J	**ICC**	**Naive OLS**	**Cluster-Robust (CR2)**	**Mixed Model (REML)**
Between-Subject	10	0.10	0.099 [0.007]	0.050 [0.005]	0.044 [0.005]
Between-Subject	10	0.50	0.280 [0.010]	0.053 [0.005]	0.053 [0.005]
Between-Subject	30	0.10	0.102 [0.007]	0.051 [0.005]	0.050 [0.005]
Between-Subject	30	0.50	0.266 [0.010]	0.058 [0.005]	0.058 [0.005]
Between-Subject	100	0.10	0.101 [0.007]	0.055 [0.005]	0.055 [0.005]
Between-Subject	100	0.50	0.260 [0.010]	0.050 [0.005]	0.050 [0.005]
Within-Subject	10	0.10	0.034 [0.004]	0.046 [0.005]	0.045 [0.005]
Within-Subject	10	0.50	0.012 [0.002]	0.052 [0.005]	0.055 [0.005]

Covariate Type	J	ICC	Naive OLS	Cluster-Robust (CR2)	Mixed Model (REML)
Within-Subject	30	0.10	0.041 [0.004]	0.054 [0.005]	0.052 [0.005]
Within-Subject	30	0.50	0.008 [0.002]	0.056 [0.005]	0.057 [0.005]
Within-Subject	100	0.10	0.044 [0.005]	0.052 [0.005]	0.057 [0.005]
Within-Subject	100	0.50	0.006 [0.002]	0.044 [0.005]	0.043 [0.005]

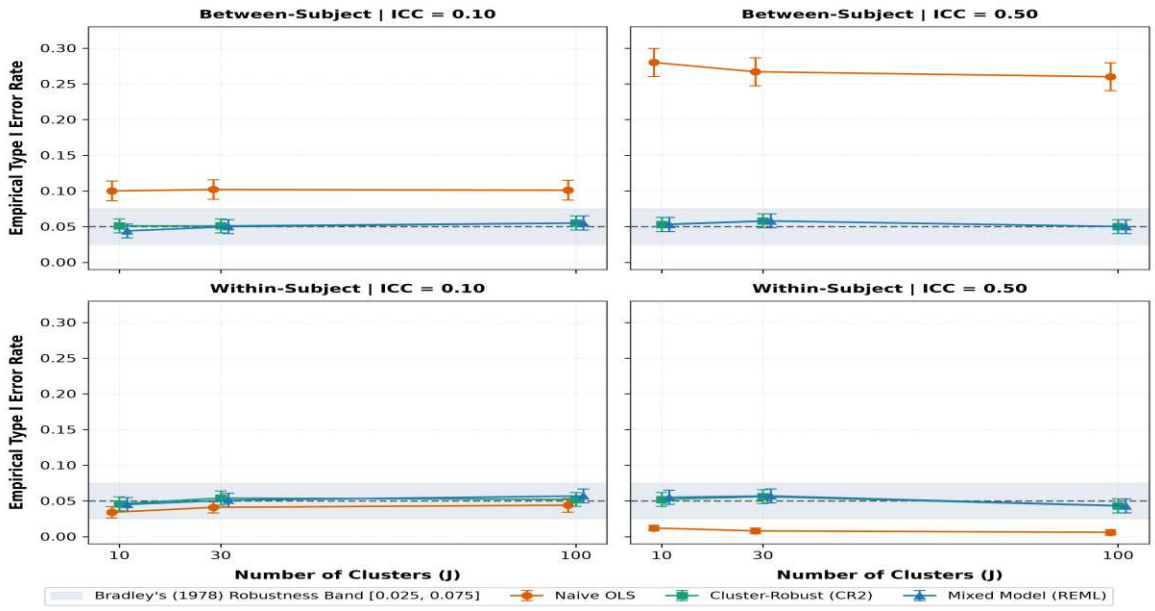


Figure 2: Empirical Type I Error Rate under the sharp null hypothesis ($\beta_1 = 0$) at nominal $\alpha = 0.05$. Shaded bands represent Bradley's (1978) liberal robustness limits $[0.025, 0.075]$*; dashed horizontal line indicates nominal* $\alpha = 0.05$.

3.5. Statistical Power and Singular Fit Audit

Table 5 summarizes statistical power under the alternative hypothesis $\beta_1 = 0.50$. For within-subject time predictors, all three methods achieved saturated 100% empirical power across all sample sizes (99.6%–100.0% at $J = 10$, and 100.0% at $J \geq 30$). Because each scenario provided $N = 5J$ observations with $\sum_{k=1}^{5} t_k^2 = 10$ (yielding a denominator sum of squares $\sum_{j=1}^{J} \sum_{k=1}^{5} t_k^2 =$

$10J$), an effect size of $\beta_1 = 0.50$ saturated the power function, masking the asymptotic efficiency penalty of naive OLS. Demonstrating empirical power loss for naive OLS on within-subject factors requires smaller effect sizes (e.g., $\beta_1 \leq 0.15$), where the wider OLS confidence intervals directly diminish detection sensitivity.

For between-subject covariates, naive OLS appeared to exhibit higher power (e.g., 49.55% at $J = 10$,ICC $= 0.50$), but this is an uninterpretable artifact of its 28.0% false-positive rate. Properly calibrated power (achieved by LMM and CR2) under strong clustering (ICC $= 0.50$) was 15.30% at $J = 10$, increasing to 40.20% at $J = 30$, and reaching 89.65% at $J = 100$. Under weak clustering (ICC $= 0.10$), properly calibrated power was substantially higher: 25.30%–26.25% at $J = 10$, 71.50%–71.65% at $J = 30$, and 99.75% at $J = 100$.

Table 5: Empirical statistical power [MCSE] across 2,000 replications (nominal $\alpha = 0.050$, true $\beta_1 = 0.500$)

Covariate Type	J	**ICC**	**Naive OLS**	**Cluster-Robust (CR2)**	**Mixed Model (REML)**
Between-Subject	10	0.10	0.426 [0.011]	0.262 [0.010]	0.253 [0.010]
Between-Subject	10	0.50	0.496 [0.011]	0.153 [0.008]	0.153 [0.008]
Between-Subject	30	0.10	0.830 [0.008]	0.716 [0.010]	0.715 [0.010]
Between-Subject	30	0.50	0.747 [0.010]	0.402 [0.011]	0.402 [0.011]
Between-Subject	100	0.10	0.999 [0.001]	0.998 [0.001]	0.998 [0.001]
Between-Subject	100	0.50	0.986 [0.003]	0.896 [0.007]	0.896 [0.007]
Within-Subject	10	0.10	0.998 [0.001]	0.996 [0.001]	0.999 [0.001]
Within-Subject	10	0.50	1.000 [0.000]	1.000 [0.000]	1.000 [0.000]
Within-Subject	30	0.10	1.000 [0.000]	1.000 [0.000]	1.000 [0.000]

Covariate Type	J	ICC	Naive OLS	Cluster-Robust (CR2)	Mixed Model (REML)
Within-Subject	30	0.50	1.000 [0.000]	1.000 [0.000]	1.000 [0.000]
Within-Subject	100	0.10	1.000 [0.000]	1.000 [0.000]	1.000 [0.000]
Within-Subject	100	0.50	1.000 [0.000]	1.000 [0.000]	1.000 [0.000]

Singular Boundary Fits in REML: Table 5b audits boundary convergence solutions (estimated random intercept variance $\hat{\sigma}_u^2 < 10^{-4}$) across all 48,000 simulated datasets. Boundary fits occurred primarily under weak clustering (ICC $= 0.10$) combined with small cluster counts ($J = 10$), reaching 24.50% for between-subject and 23.35% for within-subject designs. As J increased to 30, singular fit rates dropped to 8.95% (between) and 7.70% (within); at $J = 100$, singular fits were rare (0.75% and 0.35%). Under strong clustering (ICC $= 0.50$), boundary fits were virtually absent, occurring in only 0.65% (between) and 0.20% (within) of replications at $J = 10$, and 0.00% across all scenarios with $J \geq 30$.

Conditioning coverage on singular status reveals a fundamental structural contrast: For between-subject predictors, when REML encounters a singular boundary fit, the estimated random-intercept variance collapses to zero, and the model mistakenly relies on residual within-subject variance. Consequently, LMM coverage among singular fits drops to 91.6% at $J = 10$, ICC $= 0.10$ and 87.7% at $J = 30$,ICC $= 0.10$ (compared to 97.4% and 95.3% for non-singular fits). At $J = 10$, ICC $= 0.50$, singular fits were exceptionally rare (13 out of 2,000 replications, or 0.65%), coverage among these 13 singular cases was 69.2% (9/13), which illustrates severe variance collapse but carries substantial Monte Carlo imprecision due to the small count ($n = 13$). Because singular solutions occurred in less than 1% of replications in this cell, overall coverage remained well-calibrated at 95.2%. For within-subject predictors, because the within-subject contrast is mathematically orthogonal to the random intercept, singular fits have no adverse effect on fixed-effect coverage: LMM coverage remained 96.1% at $J = 10$ and 96.8% at $J = 30$ among singular fits.

Table 5b: REML singular boundary fit diagnostics and conditional 95% CI coverage across 2,000 replications

Covariate Type	J	ICC	Singular Fits (n/R)	Rate (%)	Overall Coverage	Singular Coverage	Non-Singular Coverage
Between-Subject	10	0.10	490 / 2,000	24.50%	0.960	0.916	0.974
Between-Subject	10	0.50	13 / 2,000	0.65%	0.952	0.692	0.953
Between-Subject	30	0.10	179 / 2,000	8.95%	0.946	0.877	0.953
Between-Subject	30	0.50	0 / 2,000	0.00%	0.956	—	0.956
Between-Subject	100	0.10	15 / 2,000	0.75%	0.954	1.000	0.954
Between-Subject	100	0.50	0 / 2,000	0.00%	0.950	—	0.950
Within-Subject	10	0.10	467 / 2,000	23.35%	0.946	0.961	0.941
Within-Subject	10	0.50	4 / 2,000	0.20%	0.950	1.000	0.949
Within-Subject	30	0.10	154 / 2,000	7.70%	0.949	0.968	0.947
Within-Subject	30	0.50	0 / 2,000	0.00%	0.958	—	0.958
Within-Subject	100	0.10	7 / 2,000	0.35%	0.944	1.000	0.943
Within-Subject	100	0.50	0 / 2,000	0.00%	0.958	—	0.958

4. Empirical Benchmark: The Sleep Deprivation Case Study

To examine how these theoretical and simulation dynamics function in real-world biomedical data, we re-analyzed the canonical sleep deprivation experiment conducted by Belenky et al. (2003) and benchmarked by Bates et al. (2015). The dataset tracks average reaction times (ms) on a psychomotor vigilance task across $N = 180$ measurements obtained from $J = 18$ subjects across 10 consecutive observation days (Days 0 to 9, where $K = 10$). In the standard `lme4` benchmarking protocol (Bates et al., 2015; Belenky et al., 2003), Day 0 represents baseline performance under normal sleep, with sleep restriction (restricted to 3 hours time-in-bed per night) commencing that night and continuing through Day 9.

We fitted four competing specifications modeling reaction time as a linear function of sleep restriction days:

1. Naive OLS Pooling: Regressing reaction time on days without clustering adjustments ($df = 178$).

2. Cluster-Robust CR2: OLS point estimation with small-sample bias-reduced standard errors clustered by subject ($df = 17$).

3. Random-Intercept LMM: LMM with subject-specific random intercepts, assuming uniform deterioration across subjects ($df = 161$).

4. Random-Slope & Intercept LMM: Maximal LMM allowing both baseline reaction times and degradation rates to vary across subjects ($df = 17$).

Table 6: Empirical comparison on the sleep deprivation benchmark dataset ($N = 180, J = 18$)

Model Specification	**Estimate (β_1)**	**Std. Error**	df	p**-value**	**95% Confidence Interval**
1. Naive OLS (complete pooling)	10.467	1.238	178	< 0.001	[8.024, 12.911]
2. Cluster-Robust OLS (CR2)	10.467	1.546	17	< 0.001	[7.206, 13.729]
3. LMM (random intercept only)	10.467	0.804	161	< 0.001	[8.879, 12.055]
4. LMM (random slope & intercept)	10.467	1.546	17	< 0.001	[7.206, 13.729]

4.1. Empirical Insights

Table 6 delivers three critical methodological insights:

The Danger of Random-Intercept-Only Models: When fitting a Random-Intercept LMM, the standard error for Days collapses to 0.804 ($df = 161$, 95% CI: [8.879, 12.055]). The model assumes that every individual deteriorates at the exact same rate (10.467 ms/day). By restricting individual slope variability to zero, the model mistakenly pools subject-by-treatment interactions into residual precision, yielding dangerously overconfident inferences (Barr et al., 2013).

Subject Slope Heterogeneity: Allowing random slopes across individuals (Model 4) increases the estimated standard error from 0.804 to 1.546 ($df = 17$, 95% CI: [7.206, 13.729]), reflecting substantial subject-level variance in neurobehavioral vulnerability to sleep restriction (Belenky et al., 2003).

Analytical Mechanism of the CR2 Estimator: Without requiring explicit specification or optimization of a random-slope covariance matrix, the Cluster-Robust CR2 estimator yielded an identical standard error of 1.546 ($df = 17$, 95% CI: [7.206, 13.729]). This exact agreement is expected in a balanced design with a single within-subject covariate, because both approaches effectively base the standard error on the between-subject dispersion of the individual OLS slope estimates. This demonstrates that CR2 reproduces the maximal random-slope standard error without requiring random-slope specification. However, as noted in Section 5.3, a single empirical dataset does not constitute universal proof of robustness to arbitrary covariance misspecification; evaluating general robustness requires Monte Carlo simulations with random slopes explicitly embedded in the data-generating mechanism (Pustejovsky & Tipton, 2018).

5. Discussion

5.1. Methodological Implications

This study demonstrates the inferential hazards of applying naive OLS regression to repeated measures data. The direction and magnitude of inferential distortion align precisely with modern design-effect formulations (Huang, 2022; Moulton, 1990). For between-subject predictors (e.g., parallel randomized trials), naive OLS standard errors underestimate true sampling variance by over 40%, causing 95% confidence interval coverage to collapse to 71.8%–74.2% and false-positive rates to surge to 26.0%–28.0%. For within-subject predictors (e.g., longitudinal trajectories), naive OLS standard errors are overly conservative (SE Ratio ≈ 1.39–1.47), producing excessively wide intervals (99.2%–99.4% coverage) and sacrificing estimation precision.

Both Linear Mixed Models (REML) with Satterthwaite corrections and Cluster-Robust CR2 estimators fully resolve these inferential breakdowns, maintaining nominal coverage (94.4%–96.0%) and controlled Type I error rates (4.3%–5.8%) down to small cluster counts ($J = 10$) (Bradley, 1978; Pustejovsky & Tipton, 2018, 2022).

5.2. Practical Decision Framework

Based on our empirical and simulation findings, we recommend the following analytical decision pathway for repeated measures designs:

Table 7: Analytical decision pathway for repeated measures designs

Cluster Count	**Analytical Goal**	**Recommended Approach**	**Requirements and Cautions**
$J < 30$ (Small)	Subject trajectories / variance components	Linear mixed model (REML)	Satterthwaite degrees of freedom is mandatory (Kuznetsova et al., 2017); audit singular boundary fits (Bates et al., 2015)
$J < 30$ (Small)	Population-averaged effects / robustness	Cluster-robust OLS (CR2)	Satterthwaite t reference distribution is mandatory (Pustejovsky, 2023); protected against random-slope misspecification in balanced designs (Pustejovsky & Tipton, 2018)
$J \geq 30$ (Moderate to Large)	Subject trajectories	Linear mixed model (REML / ML)	Asymptotically robust inference is valid (Verbeke & Molenberghs, 2000)
$J \geq 30$ (Moderate to Large)	Population-averaged effects	GEE / CR0 / CR1	Standard sandwich errors are valid; recommendation drawn from established literature (Cameron & Miller, 2015; Zeileis et al., 2020)

Note: Recommendations for GEE, CR0, and CR1 at $J \geq 30$ stem from established methodological literature (Cameron & Miller, 2015; Zeileis et al., 2020), as the present simulation focused specifically on contrasting Naive OLS, CR2, and LMM.

5.3. Limitations and Future Directions

Several methodological boundaries of the present study warrant consideration:

1. Design Regularity: The Monte Carlo simulation evaluated strictly balanced designs ($K = 5$ occasions per cluster) without missing data. In observational cohorts with variable cluster sizes (K_i) and informative attrition, CR2 degrees of freedom can decrease substantially, occasionally leading to reduced power.

2. Random-Slope Misspecification: Our simulation data-generating mechanism specified a pure random-intercept architecture. Consequently, the empirical equivalence observed between CR2 and maximal LMM in Section 4 rests on the balanced structure of the sleepstudy dataset and literature demonstrations (Pustejovsky & Tipton, 2018), rather than a simulated random-slope DGM. Future simulation work should formally manipulate random-slope variance ($\sigma^2_{u_1}$) to quantify the exact point at which random-intercept LMMs fail under controlled conditions.

3. Distributional Assumptions: All random effects and residuals were generated from Gaussian distributions. While LMM fixed-effect inference is moderately robust to non-normal random effects (Schielzeth et al., 2020), extreme heavy-tailed skewness may alter small-sample performance.

6. Conclusion

Ignoring clustering in repeated measures data creates severe inferential distortions whose direction depends strictly on covariate architecture. For between-subject factors, naive OLS generates deflated standard errors, artificially narrow confidence intervals, and catastrophic Type I error inflation (exceeding 26%), and must never be applied. For within-subject factors, naive OLS does not inflate Type I error rates, but overestimates standard errors by 39% to 47%, resulting in unnecessarily wide confidence intervals and loss of inferential efficiency.

When cluster counts are small ($J < 30$), applied researchers must adopt either Linear Mixed Models with Satterthwaite degrees of freedom or Cluster-Robust Variance Estimation using the CR2 bias-reduced linearization. For population-averaged questions, CR2 provides a transparent, assumption-lean framework that avoids the computational complexities and boundary convergence pitfalls of random-effects modeling.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest / Competing Interests: The author declares no conflicts of interest.

Data and Code Availability: Complete reproducible R and Python code pipelines, simulation routines, and benchmark dataset analysis scripts are openly accessible on GitHub: https://github.com/IsaacMoon1/repeated-measures-simulation.

Authors Contributions:

Adeyemi I.T. conceived the study, developed the methodology, executed the Monte Carlo simulations, analyzed the empirical benchmark data, and wrote the original draft of the manuscript.

Bamidele D.E. supervised the work, contributed to the methodology, and reviewed and edited the manuscript.

Adisa Z.O. and Ademola O.R. contributed to validation of the results and reviewed and edited the manuscript.

All authors read and approved the final version of the manuscript.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., & Balkin, T. J. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of Sleep Research*, 12(1), 1–12. https://doi.org/10.1046/j.1365-2869.2003.00337.x

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2), 169–181. https://www150.statcan.gc.ca/n1/pub/12-001-x/2002002/article/9058-eng.pdf

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198524847.001.0001

Huang, F. L. (2022). Using cluster-robust standard errors when analyzing group-randomized trials with few clusters. *Behavior Research Methods*, 54(1), 282–298. https://doi.org/10.3758/s13428-021-01627-0

Kahan, B. C., Forbes, G., Ali, Y., Jairath, V., Bremner, S., Harhay, M. O., Hooper, R., Wright, N., Eldridge, S. M., & Leyrat, C. (2016). Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: A review, reanalysis, and simulation study. *Trials*, 17(1), 438. https://doi.org/10.1186/s13063-016-1571-2

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3), 983–997. https://doi.org/10.2307/2533558

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963–974. https://doi.org/10.2307/2529876

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrics*, 42(1), 121–130. https://doi.org/10.2307/2531248

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 3139. https://doi.org/10.21105/joss.03139

McNeish, D., & Harring, J. R. (2017). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics - Simulation and Computation*, 46(2), 859–869. https://doi.org/10.1080/03610918.2014.983648

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

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, 72(2), 334–338. https://doi.org/10.2307/2109724

Pustejovsky, J. E. (2023). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections (R package version 0.5.10). Comprehensive R Archive Network. https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 23(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6), 110–114. https://doi.org/10.2307/3002019

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer. https://doi.org/10.1007/978-1-4419-0300-6

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, 95(1), 1–36. https://doi.org/10.18637/jss.v095.i01

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- tablesingularfits.csv