Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 5

3-6-2019

A Strategy for Using Bias and RMSE as Outcomes in Monte Carlo Studies in Statistics

Michael Harwell
University of Minnesota - Twin Cities, harwe001@umn.edu

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2938
doi: 10.22237/jmasm/1551907966

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

A Strategy for Using Bias and RMSE as Outcomes in Monte Carlo Studies in Statistics

Michael Harwell
University of Minnesota Twin Cities
Minneapolis, MN

To help ensure important patterns of bias and accuracy are detected in Monte Carlo studies in statistics this paper proposes conditioning bias and root mean square error (RMSE) measures on estimated Type I and Type II error rates. A small Monte Carlo study is used to illustrate this argument.

Keywords: Monte Carlo, bias, RMSE, analysis of outcomes

Introduction

Monte Carlo studies often focus on the impact of factors such as data distribution and sample size on a variety of outcome variables characterizing the behavior of estimators, statistical tests, and other statistical procedures such as parameter estimation algorithms. A survey of Monte Carlo studies reported 44.1%, 33.1%, 16%, and 16.8% presented results for the outcomes root mean square error (RMSE) which is used to assess bias and estimation accuracy, average bias, Type I error rate, and power, respectively. Outcomes such as model convergence rate (Depaoli, 2012) and the percentage of adequately fitting models (Beauducel & Wittmann, 2010) appear less frequently. Estimation of Type I and power rates is consistent across Monte Carlo studies but slightly different measures of bias and RMSE appear in this literature.

A standard feature of Monte Carlo studies, outcomes like RMSE, bias, Type I error rate, and power are examined separately. A strategy is presented here that

doi: 10.22237/jmasm/1551907966 | Accepted: June 21, 2018; Published: March 6, 2019.
Correspondence: Michael Harwell, harwe001@umn.edu
Michael Harwell is a Professor in the Department of Educational Psychology at the University of Minnesota.

conditions outcomes on Type I and Type II error rates to provide additional insight into patterns of bias and accuracy. This strategy also speaks to the reproducibility of substantive research findings. Stodden (2015) highlighted the important role Monte Carlo studies in statistics play in increasing the reproducibility of research findings by recommending estimators, tests, and other statistical procedures identified as possessing superior properties. Ensuring that important patterns of bias and accuracy are detected and reflected in recommendations increases the likelihood of reproducibility.

Bias and RMSE Outcomes in Monte Carlo Studies

677 articles in six journals appearing between 1985-2012 that reported Monte Carlo results in statistics were reviewed. Bias of an estimator in 210 studies (33.1%) was defined as $\left(\hat{\theta}_i - \theta\right)$, where $\hat{\theta}_i$ is an estimate of a parameter θ for the i^{th} replication (i = 1, 2,…, R). For example, $\hat{\theta}_i$ could represent a regression coefficient, standard error, or a variance component. In statistical theory the bias of an estimator is the difference between an estimator's expected value and the true value of the parameter being estimated $\left(\mathrm{E}\left[\hat{\theta}_i - \theta\right]\right)$ (Neter, Kutner, Nachtsheim, & Wasserman, 1996). Averaging $\left(\hat{\theta}_i - \theta\right)$ across R replications provides measures satisfying the definition of bias i.e.,

$$\bar{\hat{\theta}} - \theta, \quad \bar{\hat{\theta}} = \sum_{i=1}^{R} \frac{\hat{\theta}_i}{R}.$$

Although different bias measures provide slightly different information, all agree that values closer to zero show less bias.

Common bias measures include average bias:

$$\mathrm{AB} = \sum_{i=1}^{R} \frac{\left(\hat{\theta}_i - \theta\right)}{R}$$

(e.g., Finch & French, 2015); average absolute bias:

$$\mathrm{AAB} = \sum_{i=1}^{R} \frac{\left|\hat{\theta}_i - \theta\right|}{R}$$

(e.g., Yuan, Tong, & Zhang, 2015); average relative bias:

$$\text{ARB} = \left[\frac{1}{R} \sum_{i=1}^{R} \left(\frac{\hat{\theta} - \theta_i}{\theta} \right) \right] \times 100 ,$$

which is expressed as a percentage (e.g., Ye & Daniel, 2017) that can exceed 100%; and average absolute relative bias:

$$\text{AARB} = \left(\frac{1}{R} \sum_{i=1}^{R} \left| \frac{\hat{\theta} - \theta_i}{\theta} \right| \right) \times 100$$

(e.g., Culpepper & Aguinis, 2011), which can also exceed 100%. The ARB and AARB measures cannot be used if $\theta = 0$.

The AAB and AARB measures collapse under- and over-estimation and represent measures of relative error which assess bias relative to the parameter being estimated. The AB and ARB measures capture the direction of mis-estimation in the θ metric and represent measures of absolute error which assess bias as a simple difference. Expressing ARB and AARB as a percentage is helpful for interpreting the magnitude of bias but guidelines for values indicating significant bias are informal. For example, Curran, West and Finch (1996) cited Kaplan (1989) in treating ARB > 10% for chi-square statistics as indicating significant bias; Hoogland and Boomsa (1998) treated ARB > 5% as biased for factor loadings and ARB > 10% as biased for standard errors, as did Kim, Joo, Lee, Wang, and Stark (2016) for factor loadings; Jin, Luo, and Yang-Wallentin (2016) treated ARB > 5% as biased for factor loadings, and Bai and Poon (2009) treated AARB > 2.5% for slopes as showing significant bias and AARB > 5% for standard errors. Guidelines for characterizing AB and AAB values as showing evidence of significant bias are unique to individual Monte Carlo studies (e.g., Yuan et al., 2015).

It was also found that 298 studies (44.1%) reported RMSE (or RMSD, its square root), which represents the variance (or standard deviation) of the deviation of estimates about a parameter, with smaller values treated as indicating more accurate estimation (Yuan et al., 2015). Common versions of RMSE include

$$\text{RMSE}\left(\text{AB}\right) = \sum_{i=1}^{R} \frac{\left(\hat{\theta}_i - \theta\right)^2}{R}$$

(e.g., Moeyaert, Rindskopf, Onghena, & Van den Noortgate, 2017) and

$$\text{RMSE}\left(\text{ARB}\right) = \frac{1}{R}\sum_{i=1}^{R}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)^2$$

(e.g., Jin et al., 2016). A related measure of variability of estimates is

$$\text{SampVar} = \sum_{i=1}^{R} \frac{\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2}{R},$$

which estimates the sampling variance of $\hat{\theta}_i$ with larger values indicating less accurate estimation (e.g., Kohli & Harring, 2013).

An important representation of RMSE was provided by Gifford and Swaminathan (1990), who showed RMSE(AB) could be partitioned into

$$\sum_{i=1}^{R} \frac{\left(\hat{\theta}_i - \theta\right)^2}{R} = \left(\bar{\hat{\theta}} - \theta\right)^2 + \sum_{i=1}^{R} \frac{\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2}{R}, \quad (1)$$

where $\left(\bar{\hat{\theta}} - \theta\right)^2$ represents squared bias and

$$\sum_{i=1}^{R} \frac{\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2}{R}$$

represents SampVar; a similar partition exists for RMSE(ARB). This partitioning makes it possible to characterize the contributions of squared bias and sampling variance to RMSE and links RMSE to AB since $\left(\bar{\hat{\theta}} - \theta\right)^2$ is equal to $(\text{AB})^2$. If $\left(\bar{\hat{\theta}} - \theta\right)^2 = 0$ then AB = 0 and all variation among the $\hat{\theta}_i$ is due to SampVar; if

$\left(\bar{\hat{\theta}} - \theta\right)^2 = \text{RMSE}$, the $\hat{\theta}_i$ are identical, all variation among estimates is captured by squared bias, and RMSE = $(\text{AB})^2$; if $\left[0 < \left(\bar{\hat{\theta}} - \theta\right)^2 < \text{RMSE}\right]$ then AB ≠ 0 with larger $\left(\bar{\hat{\theta}} - \theta\right)^2$ values linked to greater bias.

The information provided in equation (1) suggests SampVar should be reported when possible to ensure RMSE is not misinterpreted. Equation (1) seems to be widely known (e.g., Aydin & Şenoğlu, 2015; Bray, Lanza, & Tan, 2015) although studies describing RMSE solely as a measure of accuracy still appear (e.g., Loh, Arasan, Midi, & Abu Bakar, 2017; Tofighi, MacKinnon, & Yoon, 2012). Guidelines for treating RMSE as unacceptably large are informal. For example, Hoogland and Boomsa (1998) specified RMSE(ARB) > 5% as reflecting significant bias and Bai and Poon (2009) used RMSE(ARB) > 2.5%.

Conditioning Bias and RMSE Outcomes on Type I Error Rates

An important premise is that additional insight into patterns of bias and accuracy can be obtained by conditioning measures of bias and RMSE on estimated Type I $(\hat{\alpha})$ and Type II $(\hat{\beta})$ error rates. For the Type I error case the *R* distribution of bias values can be partitioned into $R\hat{\alpha}$ and $R(1-\hat{\alpha})$ distributions, and $R\hat{\beta}$ and $R(1-\hat{\beta})$ distributions for the Type II error case. In the Type I error case average bias is computed separately for the $R\hat{\alpha}$ bias values, which are linked to statistically significant results, and the $R(1-\hat{\alpha})$ bias values, which are linked to nonsignificant results. In the Type II error case average bias is similarly computed separately for the $R(1-\hat{\beta})$ and $R\hat{\beta}$ distributions, which are linked to statistically significant and nonsignificant results. The same logic applies to conditioning on Type I and Type II error rates and only one (Type I error rates) is illustrated.

The argument for conditioning on Type I error rates is simple: Computing average bias and RMSE across *R* replications can mask important patterns and lead to potentially incorrect inferences about the properties of an estimator, test, or other statistical procedure unless the $R\hat{\alpha}$ and $R(1-\hat{\alpha})$ distributions are similar to *R*. However, there is reason to expect these distributions to often differ, in large part due to the $R\hat{\alpha}$ distribution showing more pronounced bias and poorer accuracy. A plot of the *R*, $R\hat{\alpha}$, and $R(1-\hat{\alpha})$ distributions provides insight into important patterns, and computing summary measures for each should clarify similarities and

discrepancies. Similarities among the R, $R\hat{\alpha}$, and $R(1-\hat{\alpha})$ distributions suggest reporting average bias and RMSE measures based on R is appropriate, whereas discrepancies raise questions about doing so. For example, a common pattern would be $\text{AB}[R(1-\hat{\alpha})] \leq \text{AB}[R] \leq \text{AB}[R\hat{\alpha}]$, where $\text{AB}[R(1-\hat{\alpha})]$ represents average bias computed for the $R(1-\hat{\alpha})$ distribution, AB[R] represents average bias computed for R, and $\text{AB}[R\hat{\alpha}]$ the average bias computed for the $R\hat{\alpha}$ distribution. Similarly, $\text{RMSE}[R\hat{\alpha}] > \text{RMSE}[R(1-\hat{\alpha})]$,$\text{RMSE}[R]$ is particularly likely as average bias increases. Whether differences among the distributions are sufficiently large to conclude these measures are misleading requires critical judgment or can be ignored, and it is important to acknowledge that methodological researchers may reach different conclusions. Because $R(1-\hat{\alpha})$ is a function of the R and $R\hat{\alpha}$ distributions the focus from hereon is on the latter two distributions.

Consider the Monte Carlo study of Algina and Keselman (2004). The goal was to assess the impact of three missing data conditions on five statistical procedures in a longitudinal two-group randomized trials design in which the difference in group slopes served as the estimated treatment effect. Algina and Keselman defined bias using AB with $\hat{\theta}_i$ representing the difference in group slopes and θ the true treatment effect. The outcomes included AB, sampling variance of $\hat{\theta}_i$ (SampVar), and estimated Type I error rate $(\hat{\alpha})$. Based on these outcomes the authors recommended a procedure due to Overall, Ahn, Shivakumar, and Kalburgi (1999) (OPMAOC).

As an example, Algina and Keselman (2004) reported for the three missing data conditions studied, sample size of n = 100, and R = 1,000 that AB = −.016 (SampVar = 3.47), −.035 (3.70), and .056 (3.55), respectively, for the OPMAOC procedure with estimated Type I error rates of .039, .044, and .038 (α = .05, true treatment effect = 0). The AB values indicate that in two of the missing data conditions the average treatment effect was underestimated and in a third was overestimated, whereas the SampVar (or RMSE) values suggest this parameter was estimated with similar accuracy across missing data conditions (RMSE ≈ SampVar based on equation (1)). Algina and Keselman did not provide specific guidelines for interpreting bias but their comments suggested the AB values were small. However, it's possible these measures are masking important patterns.

Table 1 outlines four possible patterns of results and conclusions based on conditioning measures of bias and RMSE on Type I errors for the Algina and Keselman (2004) study for the OPMAOC procedure, n = 100, and the third missing data condition. If $\text{AB}[R\hat{\alpha}]$ and $\text{RMSE}[R\hat{\alpha}]$ were near .056 and 3.55 (Case 1 in

Table 1), the conclusion is that AB[R] = .056 is not masking important bias patterns and estimation accuracy is adequately captured by RMSE[R] = 3.55. If $\text{AB}[R\hat{\alpha}] \approx .056$ but $\text{RMSE}[R\hat{\alpha}]$ was 28.5 (Case 2), estimation accuracy for $R\hat{\alpha}$ is eight times poorer than that for R, implying that the accuracy with which θ is estimated is less than suggested by RMSE[R] = 3.55.

If $\text{AB}[R\hat{\alpha}]$ and $\text{RMSE}[R\hat{\alpha}]$ were .56 and 3.55 (Case 3) the conclusion is that bias linked to estimating θ is greater than .056 and the accuracy with which θ is estimated is about 3.55 (RMSE[R] = 3.55, $\text{RMSE}[R\hat{\alpha}] = 3.85$). If $\text{AB}[R\hat{\alpha}] = .56$ and $\text{RMSE}[R\hat{\alpha}] = 28.5$ (Case 4) the conclusion is that AB[R] and RMSE[R] are potentially misleading, i.e. average bias appears to be greater than .056 and θ is less accurately estimated than suggested by RMSE[R] = 3.55.

Table 1. Possible bias results and conclusions for the estimated treatment effect $\hat{\theta}$ after conditioning on Type I error rate for the Algina and Keselman (2004) study for n = 100, AB = .056, SampVar = 3.55, $\hat{\alpha} = .038$, and R = 1,000

Conditioning on Type I error rate	Result	Conclusion
Case 1 $\text{AB}[R] \approx \text{AB}[R\hat{\alpha}] \approx .056$; $\text{RMSE}[R] \approx \text{RMSE}[R\hat{\alpha}] \approx 3.55$	Average bias is similar across both bias distributions. Bias contributes negligibly to RMSE and accuracy is similar across distributions.	Average bias when estimating θ is .056 and the accuracy with which θ is estimated is 3.55.
Case 2 $\text{AB}[R] \approx \text{AB}[R\hat{\alpha}] \approx .056$; $\text{RMSE}[R] = 3.55, \text{RMSE}[R\hat{\alpha}] = 28.5$	Average bias is similar across both bias distributions. Bias contributes negligibly to RMSE, but accuracy differs across distributions.	Average bias when estimating θ is .056 and the accuracy with which θ is estimated is less than suggested by 3.55.
Case 3 $\text{AB}[R] = .056, \text{AB}[R\hat{\alpha}] = .56$; $\text{RMSE}[R] = \text{RMSE}[R\hat{\alpha}] = 3.55$	Average bias based on R may be masking important patterns and contributes differentially to RMSE values. Accuracy is similar across distributions.	Average bias when estimating θ is greater than .056 and the accuracy with which θ is estimated is 3.55.
Case 4 $\text{AB}[R] = .056, \text{AB}[R\hat{\alpha}] = .56$; $\text{RMSE}[R] = 3.55, \text{RMSE}[R\hat{\alpha}] = 28.5$	Average bias based on R may be masking important patterns and contributes differentially to RMSE values. Accuracy differs across distributions.	Average bias when estimating θ is greater than .056 and the accuracy with which θ is estimated is less than suggested by 3.55.

Note: AB = average bias, SampVar = sampling variance of $\hat{\theta}$ estimates, RMSE = root mean square error of AB values = SampVar + $(\text{AB})^2$, R = number of replications, $\hat{\alpha}$ = estimated Type I error rate, $R\hat{\alpha}$ = replications producing statistically significant results.

The strategy of conditioning outcomes on Type I error rates may have particular value as $\hat{\alpha}$ departs from α (e.g., .038 vs .05). Suppose the Algina and Keselman (2004) Type I error rate of $\hat{\alpha} = .038$ for the test of the treatment effect in the above example was used to condition bias calculations. This value means 38 of R = 1,000 statistical hypotheses were incorrectly rejected and 962 were correctly retained. Suppose also $\text{AB}[R\hat{\alpha}] \approx .056$ but $\text{RMSE}[R\hat{\alpha}]$ was eight times larger than RMSE[R]. This pattern may help to explain the conservative Type I error rate of $\hat{\alpha} = .038$ because $\text{RMSE}[R\hat{\alpha}]$ (relative to RMSE[R]) increases the standard error used in testing for a treatment effect.

Methodology

An Example Using Simulated Data

To further illustrate the above arguments a small Monte Carlo study was done for the one-way random effects (two-level) model assuming continuous cross-sectional data. The underlying model was $Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$, where Y_{ij} is the score of the i^{th} level 1 unit nested within the j^{th} level 2 unit (cluster, j = 1, 2,…, J), γ_{00} is a grand mean, u_{0j} is a residual for the j^{th} cluster, and e_{ij} is a level 1 residual (Raudenbush & Bryk, 2002). In the standard model $u_{0j} \sim [\text{N}(0, \tau_{00})]$ and $e_{ij} \sim [\text{N}(0, \sigma^2)]$, where τ_{00} is the variance of cluster residuals and σ^2 is the variance of level 1 residuals.

It was shown in previous Monte Carlo studies estimates of γ_{00} generally show little bias except for small numbers of clusters (J), but the literature disagrees on the value of J needed to produce unbiased estimates of τ_{00} (Browne & Draper, 2000; Delpish, 2006; Maas & Hox, 2005). A factorial design was adopted with the factors number of clusters (J = 5, 10, 15, 20, 30) and within-cluster sample sizes (n_j = 10, 30), which were equal across clusters. In all cases model residuals were normally-distributed and homoscedastic. All programming was done in Fortran 95 and the Box and Muller (1958) method for simulating normal deviates was employed. The resulting Y variable was scaled to have a mean of 10 and variance of one.

The factorial design produced 5 (J) × 2 (n_j) = 10 conditions with R = 10,000 replications generated for each condition, which were used to estimate γ_{00} and τ_{00}. Outcomes for the Monte Carlo study were AB and RMSE(AB) based on R (i.e., AB[R], RMSE[R]), Type I error rates $(\hat{\alpha})$ for tests of γ_{00} and τ_{00} following Raudenbush and Bryk (2002), and $\text{AB}[R\hat{\alpha}]$ and $\text{RMSE}[R\hat{\alpha}]$. Least squares was used to estimate γ_{00} and restricted maximum likelihood to estimate τ_{00}; a t-test and chi-square test were used to test these parameters against zero (Raudenbush & Bryk,

2002). It is important to acknowledge that testing H_0: $\tau_{00} = 0$, which was performed by the HLM7 software (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011), is not an endorsement of this practice which has been criticized (Drikvandi, Verbekem Khodadai, & Partovinia, 2012).

Results

The Monte Carlo results are summarized in Table 2. Squared bias terms can be computed as $(AB)^2$, and sampling variance (SampVar) represents the difference between RMSE and $(AB)^2$. Two patterns emerge in Table 2: First, the bias and accuracy of $\hat{\gamma}_{00}$ estimates were generally similar for the R and $R\hat{\alpha}$ distributions of AB values with one exception: For $n_j = 10$ and $J = 5$, $\text{RMSE}[R\hat{\alpha}] = .167$ for the $(10,000)(\hat{\alpha} = .007) = 70$ bias values compared to RMSE[R] = .020, which suggests a potentially important difference in accuracy because $\text{SampVar}[R\hat{\alpha}] = .167 - (-.029)^2 = .166$ is more than eight times larger than SampVar[R] = .020 – $(.0002)^2$ = .020. A plot of the 10,000 bias values for this condition produced a unimodal and positively-skewed distribution with a skewness index of 2.17, whereas a plot of $R\hat{\alpha}$ produced a bimodal distribution with a skewness index of .13. These results suggest that reporting RMSE[R] may mask potentially important differences in estimation accuracy. The likely explanation for these results is that $n_j = 10$, $J = 5$ is adequate for minimizing average bias when estimating γ_{00} but produces less accurate estimates, an inference that may be missed if only RMSE[R] = .020 is computed. For the $n_j = 10$, $J = 10$ condition, $\text{SampVar}[R\hat{\alpha}] = .064 - (.037)^2 = .063$ is approximately seven times larger than SampVar[R] = .009 – $(.005)^2$ = .009, indicating the accuracy with which γ_{00} is estimated is less than suggested by .009.

A second pattern in Table 2 is that estimates of τ_{00} based on R replications appear to show nonnegligible bias and less accurate estimation for all $n_j = 10$ conditions and $n_j = 30$, $J = 5$. The $R\hat{\alpha}$ distribution contains (10,000)(.072) = 722 AB values producing significant results with $\text{AB}[R\hat{\alpha}] = .177$. Figure 1 shows the 722 AB values (left panel) for $n_j = 10$, $J = 5$ are generally larger and more variable than those for R (right panel). Both distributions in Figure 1 are positively-skewed with skewness indices of 1.67 and 2.16. The $\text{AB}[R\hat{\alpha}]$ values for the $n_j = 10$ conditions and $n_j = 30$, $J = 5$ range between .06 and .177 and are accompanied by $\text{RMSE}[R\hat{\alpha}]$s that are 12 to 84 time larger than their counterparts in RMSE[R] ($\text{SampVar}[R\hat{\alpha}]$ values are 30 to 195 times larger than their counterparts in

SampVar[R]). Again, it's likely much of the bias and many of the discrepancies between $\hat{\alpha}$ and α for these conditions occur because $J = 5$ is simply too small for the properties of unbiasedness and efficiency to emerge. The overall inference from Table 2 and Figure 1 is that reporting AB and RMSE values based on R replications for larger numbers of clusters and the larger cluster sample size is appropriate but results for smaller values may be masking potentially important patterns.

Table 2. Monte Carlo results

	n_j	J	AB[R]	RMSE[R]	$\hat{\alpha}$	AB$\left[R\hat{\alpha}\right]$	RMSE$\left[R\hat{\alpha}\right]$
γ_{00}	10	5	0.00020	0.02006	0.007	-0.02966	0.16748
	10	10	0.00053	0.00975	0.025	0.03792	0.06444
	10	15	-0.00018	0.00672	0.033	0.00489	0.04033
	10	20	0.00067	0.00504	0.035	0.01634	0.03428
	10	30	0.00022	0.00334	0.031	0.00433	0.02556
	30	5	0.00032	0.00661	0.008	0.00901	0.05758
	30	10	-0.00012	0.00331	0.023	0.00584	0.02249
	30	15	0.00055	0.00221	0.033	0.01572	0.01395
	30	20	-0.00003	0.00166	0.030	-0.00211	0.01156
	30	30	-0.00027	0.00109	0.030	-0.00573	0.00882
τ_{00}	10	5	0.05673	0.005490	0.072	0.17758	0.42209
	10	10	0.03853	0.002483	0.066	0.11655	0.19445
	10	15	0.03126	0.001580	0.067	0.08903	0.11111
	10	20	0.02078	0.001170	0.067	0.07471	0.07773
	10	30	0.02209	0.000780	0.064	0.06068	0.05429
	30	5	0.01841	0.000580	0.056	0.06282	0.06630
	30	10	0.01257	0.000260	0.055	0.03958	0.02698
	30	15	0.01011	0.000160	0.057	0.03105	0.01511
	30	20	0.00871	0.000120	0.058	0.02578	0.01086
	30	30	0.00008	0.007180	0.056	0.02024	0.00691

Note: All data were normally-distributed; n_j = within-cluster sample size; J = number of clusters; AB[R] = average bias based on R = 10,000 replications; RMSE[R] = root mean square error of AB values based on R = 10,000 replications; $\hat{\alpha}$ = estimated Type I error rate of tests of H_0: $\gamma_{00} = 0$ and H_0: $\tau_{00} = 0$ ($\alpha = .05$) computed as the (number of rejections) / 10,000; AB$\left[R\hat{\alpha}\right]$ and RMSE$\left[R\hat{\alpha}\right]$ represent average bias and RMSE values for replications producing significant results.

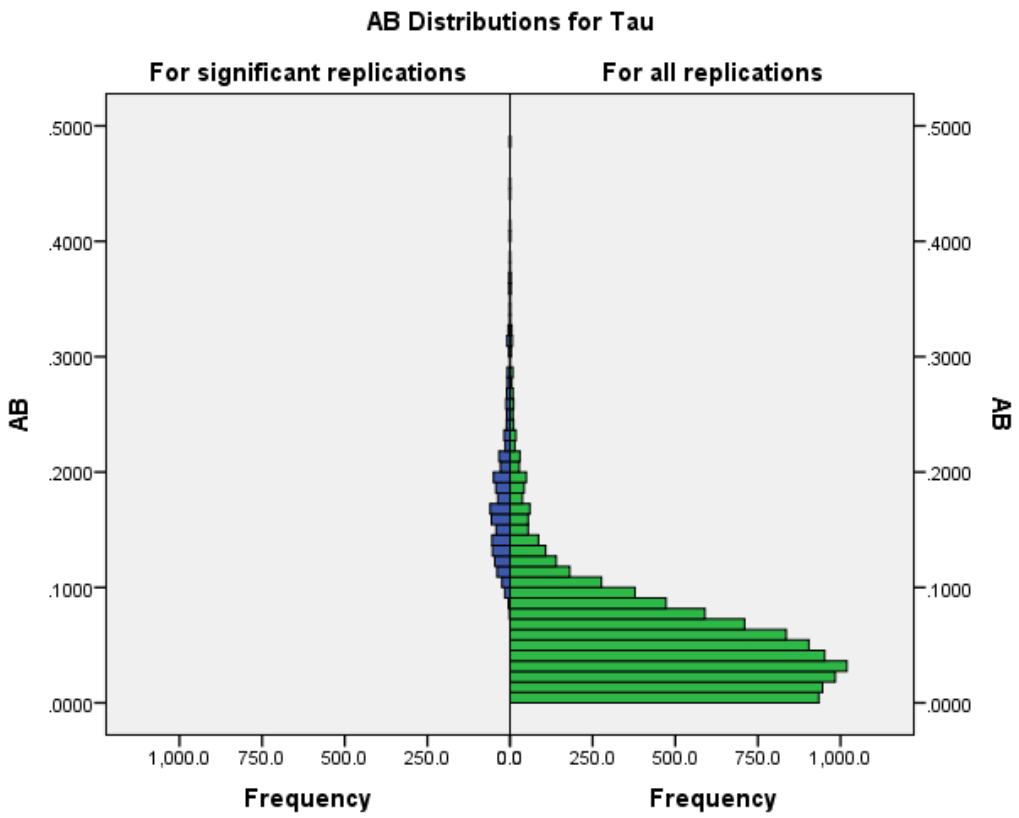

Figure 1. Distributions of AB values for estimating τ_{00} for n_j = 10, J = 5 conditioning on $R\hat{\alpha}$ (left panel) and R (right panel)

Conclusion

Monte Carlo studies in statistics investigating bias and the accuracy of parameter estimation have traditionally reported measures of average bias and RMSE based on R replications, which can mask important patterns of bias. Conditioning measures on estimated Type I error rate $(\hat{\alpha})$ provide an important complement to measures based on R in two ways: First, computing bias and RMSE for the $R\hat{\alpha}$ and $R(1-\hat{\alpha})$ distributions of bias values provides additional insight into bias patterns. In practice, examining the R and $R\hat{\alpha}$ distributions should be sufficient and if these distributions produce similar average bias values and RMSEs the inference is that reporting measures based on R is appropriate; otherwise, it's

important to evaluate the impact of results for the $R\hat{\alpha}$ distribution on study conclusions.

Second, conditioning measures of bias and RMSE on the $R\hat{\alpha}$ distribution may provide insight into the contribution of bias to estimation accuracy via equation (1), helping to clarify interpretations of RMSE. This strategy may also point to explanations for estimated Type I error rates that depart from nominal values. Conditioning evaluations of estimators, statistical tests, or other statistical procedures on Type I error rates can also enhance reproducibility by helping ensure that procedures recommended on the basis of Monte Carlo results possess superior statistical properties, which increases the likelihood of replicable findings in substantive research studies that adopt these recommendations.

The results of a small Monte Carlo study of the one-way random effects model provided empirical evidence of the value of conditioning the computation of bias and estimation accuracy on replications linked to significant and nonsignificant results. Implicit in the proposed strategy is that Type I error rates be estimated even if these are not the focus of a Monte Carlo study. Importantly, the same conditioning strategy can be used to examine patterns of bias in Type II error results.

References

Algina, J., & Keselman, H. J. (2004). A comparison of methods for longitudinal analysis with missing data. *Journal of Modern Applied Statistical Methods, 3*(1), 13-26. doi: 10.22237/jmasm/1083369780

Aydin, A., & Şenoğlu, B. (2015). Monte Carlo comparison of the parameter estimation methods for the two-parameter Gumbel distribution. *Journal of Modern Applied Statistical Methods, 14*(2), 123-140. doi: 10.22237/jmasm/1446351060

Bai, Y., & Poon, W.-Y. (2009). Using Mx to analyze cross-level effect in two-level structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(1), 163-178, doi: 10.1080/10705510802561527

Beauducel, A., & Wittmann, W. W. (2010). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(1), 41-75. doi: 10.1207/s15328007sem1201_3

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *The Annals of Mathematical Statistics, 29*(2), 10-13. doi: 10.1214/aoms/1177706645

Bray, B. C., Lanza, S. T., & Tan, X. (2015). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(1), 1-11. doi: 10.1080/10705511.2014.935265

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15*(3), 391-420. doi: 10.1007/s001800000041

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods, 16*(2), 166-178. doi: 10.1037/a0023355

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29. doi: 10.1037//1082-989x.1.1.16

Delpish, A. N. (2006). *A comparison of estimators in hierarchical linear modeling: Restricted maximum likelihood versus bootstrap via minimum norm quadratic unbiased estimators* (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

Depaoli, S. (2012). Measurement and structural model class separation in mixture CFA: ML/EM versus MCMC. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(2), 178-203. doi: 10.1080/10705511.2012.659614

Drikvand, R., Verbeke, G., Khodadadi, A., & Partovinia, V. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics, 14*(1), 144-159. doi: 10.1093/biostatistics/kxs028

Finch, W. H., & French, B. F. (2015). Modeling of nonrecursive structural equation models with categorical indicators. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(3), 416-428. doi: 10.1080/10705511.2014.937380

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14*(1), 33-43. doi: 10.1177/014662169001400104

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367. doi: 10.1177/0049124198026003003

Jin, S., Luo, H., Yang-Wallentin, F. (2016). A simulation study of polychoric instrumental variable estimation in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 680-694. doi: 10.1080/10705511.2016.1189334

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research, 24*(1), 41-57. doi: 10.1207/s15327906mbr2401_3

Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 870-877. doi: 10.1080/10705511.2016.1196108

Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research, 48*(3), 370-397. doi: 10.1080/00273171.2013.778191

Loh, Y. F., Arasan, J., Midi, H., & Abu Bakar, M. R. (2017). Inferential procedures for log logistic distribution with doubly interval censored data. *Journal of Modern Applied Statistical Methods, 16*(2), 581-603. doi: 10.22237/jmasm/1509496320

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 85-91. doi: 10.1027/1614-2241.1.3.86

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods, 22*(4), 760-778. doi: 10.1037/met0000136

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago, IL: Irwin.

Overall, J. E., Ahn, C., Shivakumar, C., & Kalburgi, Y. (1999). Problematic formulations of SAS PROC.MIXED models for repeated measurements. *Journal of Biopharmaceutical Statistics, 9*(1), 189-216. doi: 10.1081/BIP-100101008

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). HLM 7: Hierarchical linear and nonlinear modeling [computer software]. Chicago, IL: Scientific Software International.

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Applications, 2*, 1-19. doi: 10.1146/annurev-statistics-010814-020127

Tofighi, T., MacKinnon, D. P., & Yoon, M. (2012). Covariances between regression coefficient estimates in a single mediator model. *British Journal of*

Mathematical and Statistical Psychology, 62(3), 457-484. doi: 10.1348/000711008x331024

Ye, F., & Daniel, L. (2017). The impact of inappropriate modeling of cross-classified data structures on random-slope models. *Journal of Modern Applied Statistical Methods, 16*(2), 458-484. doi: 10.22237/jmasm/1509495900

Yuan, K.-H., Tong, X., & Zhang, Z. (2015). Bias and efficiency for SEM with missing data and auxiliary variables: Two-stage robust method versus two-stage ML. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(2), 178-192. doi: 10.1080/10705511.2014.935750