Correcting the Bias of the Root Mean Squared Error of Approximation under missing data

Cailey E. Fitzgerald  
University of Virginia, Charlottesville, USA

Ryne Estabrook  
Northwestern University, Evanston, USA

Daniel P. Martin  
University of Virginia, Charlottesville, USA

Andreas M. Brandmaier  
Max Planck Institute for Human Development, Berlin, Germany

Timo von Oertzen  
University of the Federal Forces, Munich, Germany, and University of Virginia, Charlottesville, Virginia

Author Note

Correspondence concerning this paper should be addressed to Timo von Oertzen, University of the Federal Forces, Werner-Heisenberg-Weg 39, 85577 Neubiberg, e-mail: timo@unibw.de.
Abstract

Missing data are ubiquitous in both small and large datasets. Missing data may come about as a result of coding or computer error, participant absences, or it may be intentional, as in planned missing designs. We discuss missing data as it relates to goodness-of-fit indices in Structural Equation Modeling (SEM), specifically the effects of missing data on the Root Mean Squared Error of Approximation (RMSEA). We use simulations to show that naive implementations of the RMSEA have a downward bias in the presence of missing data and, thus, overestimate model goodness-of-fit. Unfortunately, many state-of-the-art software packages report the biased form of RMSEA. As a consequence, the community may have been accepting a much larger fraction of models with non-acceptable model fit. We propose a bias-correction for the RMSEA based on information-theoretic considerations that take into account the expected misfit of a person with fully observed data. This results in an RMSEA which is asymptotically independent of the proportion of missing data for misspecified models. Importantly, results of the corrected RMSEA computation are identical to naive RMSEA if there are no missing data.

Keywords: missing data, structural equation modeling, fit indices, Kullback Leibler divergence, relative entropy
Correcting the Bias of the Root Mean Squared Error of Approximation under missing data

Structural Equation Modeling (SEM; Bollen, 1989) is an ubiquitous and powerful technique for the analysis of multivariate data. A host of common models can be specified and estimated within the SEM paradigm, including models based on the general linear model (GLM) like regression and ANOVA (see for example Miller, 1997), and more complex techniques ranging from factor analysis (Mulaik, 1972) non-linear growth mixture models (Grimm, Ram, & Estabrook, 2010) to general multi-level models (Curran, 2003; Bauer, 2003). Before substantial interpretations are warranted, models are usually assessed using fit indices describing how well the model fits the data. Fit indices typically compare the proposed model to either a saturated model or both a saturated and an independence model. Saturated models include all possible means, variances and covariances, and represent the best possible fit of a covariance structure to the data, while independence models include only means and variances (i.e., assume all covariances are zero) and represent the worst reasonable fit of a covariance structure. Fit indices like Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980) describe absolute misfit of a model relative to the saturated model, while indices like the Comparative Fit Index and Tucker Lewis Index (Tucker & Lewis, 1973) place the proposed model on a continuum between the saturated and independence models. 

Since its early days as a modeling framework of variances and covariances, the theory surrounding SEM has greatly expanded to include multi-level modeling (Heck, 2001), mixture modeling (Muthén & Shedden, 1999), non-linearity (Grimm & Ram, 2009), and even extensions to statistical learning approaches (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013). One important extension is the implementation of Full-Information Maximum Likelihood (FIML; cf. Finkbeiner, 1979) estimation based on raw data instead of covariance matrices. Early SEM packages relied solely on observed moment matrices (e.g., covariance matrices, mean vectors, SSCP matrices) as input, requiring researchers to convert their data into the appropriate matrices and handle missing data, ordinal data, and
other complexities prior to model fitting. Modern SEM packages can use FIML to handle raw data, eschewing user-generated data reduction in favor of fitting the model to each row of the data individually. This allows SEM to handle missing values by applying the model to whatever values are observed for a particular row of data. FIML is robust to missingness under both the missing completely at random (MCAR) and missing at random (MAR) mechanisms (Rubin, 2002; Little & Rubin, 1987), yielding unbiased parameter estimates under ignorable missingness conditions and performs better than other missing data methods, such as similar response pattern imputation or listwise deletion (Enders & Bandalos, 2001).

While the methods by which we fit SEM have changed, the methods by which we assess the fit of these models have not. Most of the common fit indices used to assess model fit make assumptions about the models being compared and the data that underly them, namely that they contain full rank covariance matrices with single values for sample size and degrees of freedom. To the extent that a particular model uses raw data under FIML and is affected by missingness, these assumptions are decidedly false and thus bias is introduced into fit indices. For example, the RMSEA quantifies the amount of noncentrality per the product of sample size and degrees of freedom. When the data are complete, the product of adjusted sample size and degrees of freedom contain one measure of the total amount of information in the sample. When some portion of the data are missing, the $\chi^2$ will decrease while the sample size and degrees of freedom values stay constant, which reduces the absolute value of the RMSEA statistic and make fit appear to be better than it would be with complete data. This artificial improvement in model fit has been noted in previous research (i.e. Davey, 2005; Hoyle, 2012) with no known solution to the problem.

Despite this, using fit statistics in the presence of missing data remains common practice. Google Scholar search indicates approximately 30% of articles (or 8,930/30,361) that cited one of two popular articles with recommended “rules of thumb” for RMSEA (Browne & Cudeck, 1993; Hu & Bentler, 1999) also include either “missing” or “full
information”. While this search is imperfect and nowhere near exhaustive, it gives some background toward the widespread use of fit statistics in SEM and, more importantly, the widespread use of fit statistics in SEM with missing data.

Summary of this article

In this article, we will first discuss the problem that missing data introduces into the realm of model fit indices in SEM, focusing on the RMSEA. We then introduce the mathematical basis for a bias correction of the RMESA. Next, we present data simulations and plots to illustrate the problem and demonstrate to what extent the proposed solutions are better suited to capture model misfit under missing data than the uncorrected RMSEA. We discuss our motivation for a correction, broader implications of the problem of missing data in SEM, limitations of our correction, and future directions in the research of model misfit indices.

Method

Mathematical Statement of the Problem

The RMSEA is an estimate based on the quantity $\sqrt{\frac{F}{df}}$, where $df$ represents the difference of degrees of freedom between a fitted model and a saturated model, and $F$ is a population quantity of the discrepancy between a normal population distribution and the best-fitting normal distribution under a given model. $F$ is thus independent of sample size and, conventionally, $\frac{\chi^2}{N}$ is used as estimator of this discrepancy. The resulting population value, $\sqrt{\frac{F}{df}}$, is thus a measure of absolute misfit per person and per degree of freedom. The fraction $\sqrt{\frac{\chi^2}{N \cdot df}}$ is an estimate of this population quantity for model misfit, which forms the basis of RMSEA.

For complete data, the RMSEA for a proposed model is defined as

$$RMSEA_{uncorrected} = \sqrt{\max \left(0, \frac{\chi^2 - df}{(N - 1)df}\right)}$$

\footnote{As of December 1, 2013.}
where $df$ is the difference in degrees of freedom between the proposed model and a saturated normal model, $N$ the number of participants, and $\chi^2$ the difference in the minus two log-likelihood between the proposed and saturated models for the given data set. Without model misspecification, the $\chi^2$ index will be distributed as a central $\chi^2$ distribution with $df$ degrees of freedom; therefore, the fraction will have an expected value of zero, with decreasing sampling variance as $N$ increases. If the model is misspecified, $\chi^2$ follows a non-central $\chi^2$ distribution with a non-centrality parameter proportional to $N$, and thus RMSEA will converge towards a fixed non-zero number in expectation. The degree to which RMSEA is larger than zero is therefore a measure of model misfit.

As a data set has an increasing proportion of missing data (for example, assuming missing completely at random; MCAR), the minus two log likelihoods of both the proposed and saturated models will in expectation decrease proportionally to that amount. In consequence, the $\chi^2$ – the difference between the two – will also decrease proportionally to the proportion of missing data, while $df$ and $N$ remain constant. Hence, RMSEA values will become smaller in value and, hence, more optimistic for increasing proportions of missing data. Note that when there is no model misspecification, the proportion of missing data does not bias RMSEA.

Defining the Correction

How can we best correct the computation of RMSEA such that it is invariant under the proportion of missing data for misspecified models? We require that the bias-correction must leave RMSEA unchanged if there are no missing data. Second, we require the bias-correction to yield asymptotically identical RMSEA values under misspecification and different levels of missing data under the assumption of MCAR and MAR. In the following, we propose a method to correct the bias incurred in the RMSEA by missing data.

To account for the downward bias in the $\chi^2$ as estimator of model misfit, the $\chi^2$ value can be replaced by an estimate of its expected value if no missing data were present. We
obtain this bias correction by computing the expected divergence of a single person with complete observations and the best-fitting model and multiply this by the effective sample size, that is, the sample size multiplied by the proportion of observed values, to obtain the expected discrepancy if there had been no missing data.

Under MCAR or MAR missingness, the model-implied covariance matrix from a model with missing data is an estimate of the population covariance matrix, and by extension, of the model-implied covariance matrix if no missing data were present. Using the Kullback-Leibler divergence (Kullback & Leibler, 1951), also known as relative entropy, between the model-implied distribution and the saturated model’s distribution, we compute the expected value of the \( \chi^2 \) value of a single, fully observed person. Hence, the negative two log-likelihood of the complete data can be estimated as \( 2N \) times the Kullback-Leibler (Kullback & Leibler, 1951) divergence between the two distributions, just as \( 2N \) is multiplied by the fit function in covariance modeling to generate \( \chi^2 \) values. \( 2KL \) is given by

\[
2KL = \text{Trace} \left( \Sigma^{-1} S \right) + (\mu - m)^T \Sigma^{-1} (\mu - m) - k - \ln \left( \frac{|S|}{|\Sigma|} \right)
\]

where \( \Sigma \) and \( \mu \) are the model distribution covariance matrices and mean, \( S \) and \( m \) the saturated model distribution covariance matrix and mean, and \( k \) the number of variables (i.e., the dimensionality of the distributions).

Note that \( 2 \cdot KL \) is identical to \( \frac{\chi^2}{N} \) for complete data sets and, thus, the bias correction will not change RMSEA for complete data sets. To ensure an unbiased estimation of the average discrepancy for missingness with rate \( 1 - p \), we need to discount the sample size, \( N \), by the proportion of observed cases, \( p \). The resulting term \( 2pN \cdot KL \) has the same expectation as \( \chi^2 \) for all rates of missingness if \( S = \Sigma \), i.e., if the model is correctly specified. So we obtain a bias correction if we replace our estimator of \( F \) with \( 2KL \). For \( S \neq \Sigma \) in the population, \( 2pN \cdot KL - df \) has a non-zero expectation that is asymptotically proportional to \( pN \). Adding the maximum operator to avoid negative
values under the square root, we obtained the bias-corrected RMSEA:

\[
RMSEA_{KL} = \sqrt{\max\left(0, \frac{2KL}{df} - \frac{1}{pN}\right)}
\]

Note that even though the fraction term is guaranteed to have a constant expectation for different missingness rates, the expression under the square root (which includes the \(\max\) operator) may show an increasing expectation with \(p\) if the distribution of the fraction term includes negative results. However, the actual center of the distribution is not changed; the effect is just due to the negative values being moved to zero. The median and mode of the distribution stay constant.

Simulation Design

To demonstrate the effect of the proposed bias correction, we repeatedly simulated data and fitted it with a misspecified model. Model specification and estimation was done in \(\Omega\)nyx (von Oertzen, Brandmaier, & Tsang, 2015) and lavaan (Rosseel, 2012). We present simulations based on variations of a simple two-variable model, and a factor model with four indicators. In the simple models used for data generation, two observed variables were created, each with variance one and covariance between them set at intervals ranging from 0.0 to 0.9 (Figure 1). The second variant of the simple model (Figure 2) is similar to the first, with two observed variables, each with freely estimated variance and a covariance constrained to a value of 0.0. The mean structure was ignored by fixing it to zero for both variables. This second model was then used to fit the simulated data. The model was chosen as the most simple constellation of two variables to isolate the effect of missing data under misspecification on the RMSEA.

In addition to generating raw data, simulation setup included a structured environment in which missing values could be simulated. All data were simulated with \(N = 1000\), and missing value ratios ranged from 0.0 to 0.75, increasing in increments of 0.05. All simulation conditions were repeated for 1000 trials.
Following data generation, two types of missing data were simulated: data missing completely at random (MCAR) and data missing at random (MAR). MCAR data were simulated by randomly deleting variables by way of introducing a random number generator into the model. Using an independent Bernoulli process, data points were set to missing according to specified probabilities between 0.0 and 0.75.

For simulations with MAR, the generating process for missing data was set up differently. Missingness under MAR depends on observed values of non-missing data, so care must be taken to not create a missing value of a variable that was previously used to induce missingness elsewhere. Thus, we induced missingness in one variable only as a function of the other variable, resulting in a range of total missingness lower than what it was for MCAR. The total missingness ratio for MAR ranged from 0.0 to 0.4. A sigmoid function was used to generate a probability of missing of the second variable that depended on the value of the first variable and ranged between 0 and 1. Where the MCAR simulation used a constant value for $p_{miss}$ across all subjects, the value of $p_{miss;i}$ for the $i$-th subject was generated using a sigmoid squashed value of the observed value $x_i$ for the first variable of that subject scaled by the overall missing rate $p_{miss}$.

$$p_{miss;i} = \frac{2 \cdot p_{miss}}{1 + e^{-x_i}}, 0.0 \leq p_{miss} \leq 0.5$$

After removing the missing observations according to the specified process, data was fitted to the second, misspecified model that had fixed zero covariance between the two variables. Uncorrected and $KL$-corrected RMSEAs were then computed for the estimation result. All simulation conditions were repeated for 1,000 trials. Data generation and calling the $\Omega$nyx SEM program was done in Java.
Results

Description of Plots

Results are presented visually by condition, with one plot per type of missingness (MCAR vs. MAR) and observed covariance between variables. The horizontal ($x$) axes of each plot represent the percent of data missing, while the vertical ($y$) axes of each plot represent RMSEA. The standard error of the values given in the plots ranged from 0.0002 to 0.002. To aid in the interpretation of these simulations, we follow general guidelines in the literature to define RMSEA values less than 0.05 to indicate good fit, greater than 0.1 to indicate poor fit, and values in between 0.05 and 0.1 to indicate mediocre fit (Browne & Cudeck, 1993). A line for RMSEA of 0.05 is included where necessary.

Simulation results revealed consistent patterns for RMSEA calculations. RMSEA values were most strongly related to model misspecification (i.e., the strength of the correlation between variables that was constrained to zero in the misspecified model), but also showed effects of missingness patterns (MCAR vs. MAR). In the bivariate model, for both covariance values of 0.0 and 0.125, MAR and MCAR data generally yielded similar patterns.

**MAR and MCAR, \( cov > 0.0 \).** For data MAR with a covariance of 0.125 between variables (Figure 3), the uncorrected RMSEA values scale linearly with missingness. For the data simulation conducted here, a missingness percentage of zero results in an RMSEA value of approximately 0.065, which indicates that the model has mediocre fit. With only 30% missingness, however, the RMSEA value drops below 0.05, indicating “good” fit. With an additional decline of 20% of data reported, 50% missingness total, we see the RMSEA value decreases by almost two one-hundredths, confirming the linearly declining pattern observed. \( KL \) corrected RMSEA values remain constant across all levels of missingness, with only a slight increase near 35% missingness.

For data MCAR with a covariance value of 0.125 between variables (Figure 4), we see a similar pattern as was observed with the MAR data condition above. In these
calculations, however, the trends are even more dramatic. Note the difference in scale between the MCAR and MAR conditions; as previously described, data that we simulated to be MAR can have missing percentages only half of what missingness percentages in data simulated as MCAR. Uncorrected RMSEA values continue to decline linearly as missingness percentage increases. With increasing percentages of missing data, uncorrected values reach the .05 cutoff rather quickly and continue to decline far below the original RMSEA value of approximately 0.065. Values for $KL$ corrected RMSEA decline and reach the 0.05 cutoff value and thus cross the threshold into the range of RMSEA values indicating “good” fit.

As model misspecification increases, the resulting patterns observed become more pronounced. In the condition with a covariance of 0.9, we observe a similar trend, with uncorrected values scaling downward as the percentage of missingness increases, and $KL$-corrected values being almost completely independent of missingness. Figures 5 and 6 display results for data simulation when model covariance $cov = 0.9$ under MAR and MCAR data conditions, respectively. The mean of the $KL$ corrected RMSEA increases slightly for high missingness values. This is due to the floor effect in which $KL$-corrected values are found using the maximum operator. This results in only positive RMSEA values being reported, skewing the mean in a positive direction. When the expression beneath the square root sign is evaluated and is thus no longer required to be positive, the floor effect is eliminated.

**MAR and MCAR, $cov = 0.0$.** For models with no model misspecification (i.e., those with no covariance between variables), we observe no decline in RMSEA values for any of the three calculations. Instead, uncorrected RMSEA values remain constant at a level of almost zero, even as missingness increase. Our proposed bias-corrected RMSEA, shows slight increases as missingness increases, that is, it has a slight tendency to overpessism in adjudging model fit. When missing percentage values reach 35-40%, both MAR and MCAR $KL$ corrected values show an increase in terms of RMSEA value. See
Figures 7 and 8 for these results.

**MCAR, factor model.** As a final, more practically relevant simulation, we set up a confirmatory factor model with four indicators in lavaan (Rosseel, 2012). The population factor loadings were 1.0, 0.7, 0.8, 0.6. Also, the population model included two additional item covariances that were not modelled incurring severe misspecification. Items 1 and 2 had a covariance of 0.6 and items 3 and 4 had a covariance of 0.6. Sampling a data set of 1,000 participants without missing data and fitting this using a misspecified factor model yields RMSEA values of approximately a range of 0.54 to 0.62 indicating severe misfit. Under MCAR, we repeatedly simulated data from the population model and used the misspecified model to compute both RMSEA and KL-corrected RMSEA. Figure 9 show the results. Each condition represents an average over 100 runs. As we have seen in the simpler models, naive RMSEA values drastically increase towards seemingly better model fit with higher proportions of missing data while bias-corrected values remain largely constant.

**Results Summary**

Among models with some misspecification, we see a a downward bias with the uncorrected RMSEA calculation and largely constant values for the $KL$ corrected RMSEA. This shows up among both MCAR and MAR data simulations among models with both low (cov = 0.125) and high (cov = 0.9) model misspecification.

For the models having no misspecification, an upward trend for $KL$ corrected values is observed when missingness values are high, but not when they are low. In this case, the uncorrected RMSEA is the best choice and the $KL$-correction introduces a slight pessimistic bias.

**Discussion**

The RMSEA, as defined by Steiger and Lind (1980) may be regarded as standardization of the $\chi^2$ index of model misfit. Under the null hypothesis of no misfit, that is the population distribution can be represented by the hypothesized model, the
RMSEA follows a $\chi^2$ distribution for any proportion of missing data (MAR and MCAR). Hence, we find in our simulation that the uncorrected RMSEA is constant in expectation for every proportion of missing data when there is no model misspecification. We conclude that, under the null hypothesis, the uncorrected RMSEA behaves correct and according to our expectations. We also find that both the uncorrected RMSEA underestimates misfit and is thus an overoptimistic measure of goodness-of-fit when missing data are present. The overoptimism may severely bias our conclusions about correct model misspecification. In our simulations, the suggested bias correction decreases this overoptimism such that in many instances, models would be rejected as not well fitting even though the uncorrected RMSEA remains below typically used cut-off criteria for “good” model fit.

RMSEA was created to detect model misspecification. In fact, if we want a test whether a model is perfectly specified (to be precise, testing and possibly rejecting the null hypothesis of perfect fit), we could directly use the likelihood ratio test. The logic of the RMSEA is that models are never perfectly specified, and in fact would be useless if they were, since simplification is an integral part of statistical modeling. The RMSEA is a quantification of the degree of misspecification that allows us to work with “slightly” misspecified models that still are deemed to work as desired in most cases. We showed that with misspecification, the uncorrected RMSEA is no longer guaranteed to be constant over differing levels of misspecification. This is evident from our simulations, where the uncorrected RMSEA decreases with higher proportions of missing data. We proposed a bias correction for the RMSEA adjusting the estimate of $F$ via the Kullback-Leibler divergence such that RMSEA remains an appropriate fit index when data is only partially observed.

We showed that the $KL$ corrected RMSEA index is largely unaffected by the proportion of missing data, and therefore provides a better indicator of model fit. We corroborated our findings by simulations.
Limitations

Although the simulations strongly support the notion that $KL$-corrected RMSEA values are unaffected by missing data, further investigation is needed for more general models. The model used in our simulation was minimal and was designed as a proof of concept that demonstrates the behavior of our proposed corrections. In most cases, such models are oversimplified and do not adequately represent more complicated datasets. Increasing the complexity of data simulation models and reassessing RMSEA behavior will allow for broader confirmation of the effectiveness of the $KL$ correction as a modification for the RMSEA fit index.

A possible objection to implementing the $KL$ correction for RMSEA lies in the realization that, in presence of missing data, corrected RMSEA values indicate worse model fit than uncorrected values. Though our rationale for an RMSEA correction is rooted in theory and practice, it may be hard to accept that RMSEA values that have been reported thus far have been implying artificially inflated goodness-of-fit for models estimated from missing data. Accepting the $KL$ correction will pressure researchers to reassess model fit, and there may be resistance if $KL$ corrected RMSEA values point to a different conclusion than uncorrected RMSEA values had indicated previously. Considering that the typically used cut-off values are based on rules-of-thumb derived from observed RMSEA values without using the correction, it may be a viable strategy to re-think the typically employed cut-off values. Changing the RMSEA calculation under missing data at large, with higher RMSEA values considered acceptable, may reduce resistance to such a change without necessarily accepting considerably worse models as acceptable but rather restore fairness among models derived from completely and partially observed data. RMSEA values may then become comparable across different levels of missingness, and preference will no longer be given to models where missing data rates are high.
Summary and Future Directions

The original formulation of RMSEA was derived under the assumption of complete observations. We found that a naive computation of the RMSEA under missing data results in overoptimistic model fit (i.e. RMSEA values that are too low). Hence, we must conclude that most reported model fits using RMSEA are too optimistic. We propose that the RMSEA can be corrected by replacing the $\chi^2$ with an estimator based on the $KL$ divergence in the RMSEA equation. This leads to an intuitive extension of the original idea behind the RMSEA to settings of missing data.

This paradigm can and should be extended to other fit indices as well, and be incorporated into the updating of SEM fit statistics to account for novel estimation approaches. Other fit indices are affected by missing data, and the $KL$ correction should be explored and applied to CFI, TLI, and other indices. As the $KL$ correction is an estimate, estimation accuracy needs to be considered when creating confidence intervals around $KL$-corrected RMSEA. Beyond these corrections, we must accept and further explore ways of evaluating SEM model fit that do not depend on the assumption of a complete saturated model, especially as we extend SEM into planned missingness, design and definition variables, and other important advances that take SEM from a covariance method to flexible GLM-based modeling.

Informing researchers about the benefits of an RMSEA correction and its ability to handle missing data may encourage the use of SEM in the realistic scenario of missing data. In order to do this, it is important to provide the scientific community easy access to the correction. A feasible solution is writing a package for R which includes wrapper tools that will extract data from existing functions and calculate RMSEA corrected values. In this line, we attach a small R script to this article that derives $KL$-corrected RMSEA values for a given estimation result. The script is available in the supplemental materials. Omnyx is currently the only software providing both uncorrected and corrected RMSEA. We hope that eventually other software developer teams will follow and integrate the $KL$
correction into existing data software tools to encourage and support widespread use of corrected RMSEA estimates.
References


Figure 1. Model used for data generation. The cov values range from 0.0 to 0.9 in the simulation.
Figure 2. Model used to fit data. The covariance value between variables 1 and 2 is fixed to zero, which is a misspecification if the data was generated with nonzero covariance.
Figure 3. Simulation results for data generated with a covariance of 0.125 and data missing at random. While $KL$ corrected RMSEA values remain mostly constant across levels of missingness, uncorrected RMSEA yield artificially improved model fit. Standard errors ranged from 0.0002 (left) to 0.002 (right).
Figure 4. Simulation results for data generated with a covariance of 0.125 and data missing completely at random. Again, $KL$ corrected RMSEA values stay mostly constant across levels of missingness, while uncorrected RMSEA yield artificially improved model fit. Standard errors ranged from 0.0002 (left) to 0.002 (right).
Figure 5. Simulation results for data generated with a covariance of 0.9 and data missing at random. Again, $KL$-corrected RMSEA values remain mostly constant across levels of missingness and uncorrected RMSEA yield artificially improved model fit. Standard errors ranged from 0.0002 (left) to 0.002 (right).
Figure 6. Simulation results for data generated with a covariance of 0.9 and data missing completely at random. Again, \( KL \) corrected RMSEA values remain mostly constant across levels of missingness and uncorrected RMSEA yield artificially improved model fit. Standard errors ranged from 0.0002 (left) to 0.002 (right).
Figure 7. Simulation results for data generated with no model misspecification and data missing at random. As expected, results stay close to zero. Corrected RMSEA indices start increasing for high missingness rates. Standard errors were around 0.0001. When evaluating RMSEA values in this condition, keep note of the small scale of the y-axis.
Figure 8. Simulation results for data generated with no model misspecification and data missing completely at random. As expected, results stay close to zero. Corrected RMSEA indices start increasing for high missingness rates. Standard errors were around 0.0001. When evaluating RMSEA values in this condition, keep note of the small scale of the y-axis.
Figure 9. Simulation results for data generated from a factor model with four items with severe model misspecification and data missing completely at random.