



Indirect Effects in Sequential Mediation Models: Evaluating Methods for Hypothesis Testing and Confidence Interval Formation

Davood Tofighi^a  and Ken Kelley^b 

^aDepartment of Psychology, University of New Mexico; ^bMendoza College of Business, University of Notre Dame

ABSTRACT

Complex mediation models, such as a two-mediator sequential model, have become more prevalent in the literature. To test an indirect effect in a two-mediator model, we conducted a large-scale Monte Carlo simulation study of the Type I error, statistical power, and confidence interval coverage rates of 10 frequentist and Bayesian confidence/credible intervals (CIs) for normally and nonnormally distributed data. The simulation included never-studied methods and conditions (e.g., Bayesian CI with flat and weakly informative prior methods, two model-based bootstrap methods, and two nonnormality conditions) as well as understudied methods (e.g., profile-likelihood, Monte Carlo with maximum likelihood standard error [MC-ML] and robust standard error [MC-Robust]). The popular BC bootstrap showed inflated Type I error rates and CI under-coverage. We recommend different methods depending on the purpose of the analysis. For testing the null hypothesis of no mediation, we recommend MC-ML, profile-likelihood, and two Bayesian methods. To report a CI, if data has a multivariate normal distribution, we recommend MC-ML, profile-likelihood, and the two Bayesian methods; otherwise, for multivariate nonnormal data we recommend the percentile bootstrap. We argue that the best method for testing hypotheses is not necessarily the best method for CI construction, which is consistent with the findings we present.

KEYWORDS

Indirect effect; confidence interval; sequential mediation; Bayesian credible interval

Theories hypothesizing and studies testing sequential mediation chains, in which two or more mediators are sequentially measured over time, have become prevalent across a variety of areas in psychology (e.g., Ato García, Vallejo Seco, & Ato Lozano, 2014; Bernier, McMahon, & Perrier, 2017; Deković, Asscher, Manders, Prins, & van der Laan, 2012; Koning, Maric, MacKinnon, & Vollebergh, 2015; Reh, Tröster, & Van Quaquebeke, 2018). We focus on the mediation model in Figure 1, which illustrates a sequential two-mediator chain. In particular, Figure 1 shows an empirical example in which there is a random assignment to drink-refusal training (X), which is hypothesized to improve resistance skills (M_1), which is then hypothesized to reduce intention to drink alcohol (M_2), which ultimately leads to reduced drinking following treatment (Y). Under a set of correct specification assumptions, including the assumption that there are no omitted variables that influence the posited variables in the mediation model and the mediators and outcome variable are continuously distributed, the magnitude of the specific indirect effect of X on Y through

M_1 and M_2 is the product of the regression (path) coefficients, $\beta_1 \times \beta_2 \times \beta_3$ (VanderWeele, 2015).

Two important outcomes of conducting a sequential mediation analysis are (a) the test of the null hypothesis of no indirect effect and (b) the confidence/credible interval (CI) for the population indirect effect. To evaluate the types of methods used to test indirect effects in sequential mediation analysis, we conducted a survey of published literature in several areas of psychology from 2017 to 2018 to investigate methods currently used and recommended (see the supplemental materials for details). In addition, we reviewed methodological journal articles and books (e.g., Falk & Biesanz, 2014; Fritz, Taylor, & MacKinnon, 2012; Hayes, 2013; MacKinnon, 2008; Preacher & Hayes, 2008; Shrout & Bolger, 2002; Taylor, MacKinnon, & Tein, 2008; Williams & MacKinnon, 2008) that advocate specific tests of indirect effects in sequential mediation analysis. Our survey identified several critical issues that have not been thoroughly addressed in the literature, concerning the test of an indirect effect in a sequential

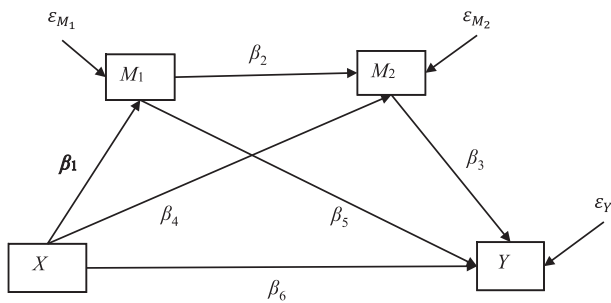


Figure 1. A two-mediator sequential mediation chain in which the mediators are sequentially related. The model has one antecedent (independent) variable, X (drink refusal training), two sequential mediators, M_1 (resistance skills) and M_2 (intention to drink alcohol), and one outcome variable, Y (number of drinks per week). Rectangles show observed variables. An arrow between two variables indicates a linear regression effect of the variable on the left, on the other variable. Term β denotes a population coefficient (path) for a linear regression in which a dependent (endogenous) variable (e.g., an outcome variable or a mediator) is predicted by another endogenous variable or an independent (exogenous) variable. Term ε denotes a residual term for each dependent (endogenous) variable. Under the no-omitted-confounder assumption, a specific indirect effect of X on Y through M_1 and M_2 equals $\beta_1 \beta_2 \beta_3$.

two-mediator model and CI formation for the population value of the indirect effect.

Among other things, our review highlighted a lack of comprehensive Monte Carlo simulation studies to evaluate six promising, but understudied methods of CI formation and for testing an indirect effect in sequential mediation model, single-mediator model, or both. These methods include two variations of Bayesian credible interval (Muthén & Asparouhov, 2012; Yuan & MacKinnon, 2009), one with *flat* priors (Bayes-Flat; noninformative, uninformative) and one with *weakly informative* (Bayes-Weak) priors for regression coefficients (note that this is the default specification in the *rstanarm* package; Muth, Oravecz, & Gabry, 2018; Stan Development Team, 2018). Additionally, we evaluate other methods from the frequentist approach, such as (a) the profile-likelihood method (Neale & Miller, 1997; Pawitan, 2001; Pek & Wu, 2015), (b) the Monte Carlo CI with maximum likelihood (ML) standard errors method (MC-ML; MacKinnon, Lockwood, & Williams, 2004; Preacher & Selig, 2012; Tofighi & MacKinnon, 2016), (c) the robust Monte Carlo (MC-Robust) CI method, which is an extension of the MC-ML,¹ (d) a semi-parametric

(model-based) bootstrap using Bollen-Stine bootstrap (BS; 1992), and (e) another semi-parametric bootstrap using the Yuan, Hayashi, and Yanagihara (YHY, 2007) method. To our knowledge, no Monte Carlo simulation study to date has examined the Bayes-Weak, BS, and YHY methods for any single-mediator or sequential mediation model. Previous Monte Carlo simulation studies have examined profile-likelihood CI, MC-Robust, and Bayes-Flat (Chen, Choi, Weiss, & Stapleton, 2014; Cheung, 2007; Falk, 2018; Falk & Biesanz, 2014) for single-mediator models; however, there is no published work for these methods in the context of a sequential mediation model.

The shape of the sampling distribution of the indirect effect in a sequential two-mediator model is different from that of the indirect effect in a single-mediator model. Recall that the indirect effect of a sequential two-mediator model is the product of three coefficients, whereas the indirect effect for a single-mediator model is the product of two coefficients. Because the performance of the methods to form an interval estimate depends on the shape of the sampling distribution of indirect effect, as well as the size of the parameters, and sample size, generalization of the statistical evaluation of these methods from a single-mediator to a sequential mediation model is premature. For example, Williams and MacKinnon (2008) concluded that the percentile, BC bootstrap, and MC-ML methods showed worse Type I error and coverage in a sequential two-mediator model than in a single-mediator model. Thus, because of the growing importance of sequential mediation, it is imperative to have a formal evaluation of the competing methods so that recommendations can be made to researchers.

In addition to the complications enumerated, the assumption of the normality of the residual terms, henceforth simply referred to as the assumption of normality, is often violated in psychological science data (Cain, Zhang, & Yuan, 2017; Micceri, 1989). When the assumption of normality is violated, ML estimates in large samples remain consistent, but are less efficient (Andreassen, Lorentzen, & Olsson, 2006; Olsson, Foss, Troye, & Howell, 2000). The standard errors for the model parameters and indirect effect estimates tend to be inconsistently estimated (Finch, West, & MacKinnon, 1997), and methods such as bias-corrected and accelerated (BCa) CI tend to show inconsistent coverage and inflated Type I error rate (Biesanz, Falk, & Savalei, 2010). Further, the likelihood-ratio test statistic might not have a chi-squared distribution for a smaller sample size, thus adversely

¹We will use the MC-Robust CI with Huber-White (Huber, 1967; White, 1980) standard errors (and robust covariance of the parameter estimates) to adjust for the potential non-normality of data; however, Falk (2018) used MC-Robust with robust Satorra-Bentler (2010) standard error correction.

impacting the performance of the chi-squared-based methods such as profile-likelihood CI and fit indices (West, Finch, & Curran, 1995). Even when the assumption of normality of residuals holds, the sampling distribution of an indirect effect is not normally distributed especially in smaller sample sizes and effect sizes (Craig, 1936; Springer & Thompson, 1966). All previous studies of sequential mediation model have evaluated the performance of tests of indirect effects with normal data (Taylor et al., 2008; Tofighi & MacKinnon, 2016; Williams & MacKinnon, 2008). For a single-mediator model, Finch et al. (1997) studied impact of various degrees of nonnormality, “moderate” (skewness = 2, kurtosis = 7) and “extreme” (skewness = 3, kurtosis = 21), on standard error and bias of indirect effect, but not on CI coverage. Biesanz et al. (2010) studied the effect of moderate nonnormality (skewness = 2, kurtosis = 7) of the outcome variable (Y), not the mediator (M), using the percentile and BCa bootstrap CI for performing a test of the null hypothesis. Using the latent mediator and outcome model, Falk (2018) studied the effect of nonnormality (skewness = 1.98, kurtosis = 9.59, which is close to what the two previous studies considered “moderate” nonnormality) of the indicators for the latent variables on the Type I error, power, and coverage of the percentile, BC, MC-ML, MC-Robust, and profile-likelihood methods. For a single-mediator model, the effect of nonnormality on the following methods has not been considered: Bayes-Weak, Bayes-Flat, BS, and YHY.

The purpose of this article is to address these critical issues and to provide a solid foundation for making recommendations to researchers when interest concerns testing the null hypothesis of no sequential mediation and reporting a CI for the population indirect effect. To begin, we review 10 recently developed and existing methods for constructing a CI and testing indirect effects: (a) Bayesian credible interval with flat prior (Bayes-Flat), (b) Bayesian credible interval with weakly informative prior (Bayes-Weak), (c) Monte Carlo (parametric bootstrap) CI with ML standard error (MC-ML; MacKinnon et al., 2004; Preacher & Selig, 2012; Tofighi & MacKinnon, 2016), (d) robust Monte Carlo CI with Huber-White (Huber, 1967; White, 1980) standard errors (MC-Robust), (e) Bollen and Stine (BS; 1992) semi-parametric (model-based) bootstrap, (f) Yuan, Hayashi, and Yanagihara (YHY, 2007) semi-parametric (model-based) bootstrap, (g) profile likelihood (Neale & Miller, 1997; Pawitan, 2001; Pek & Wu, 2015), (h) percentile non-parametric bootstrap methods (Bollen & Stine, 1990;

MacKinnon et al., 2004), (i) bias-corrected (BC) non-parametric bootstrap (Efron, 1987; MacKinnon et al., 2004; Shrout & Bolger, 2002), and (j) bias-corrected and accelerated (BCa) nonparametric bootstrap (Efron & Tibshirani, 1993). We then conduct a large-scale Monte Carlo simulation study examining the Type I error when there is no mediation, statistical power when mediation exists, and CI coverage for the 10 methods across combinations of sample sizes and values of regression coefficients for both multivariate normal and nonnormal data. We focus on the two-mediator sequential model in Figure 1. Finally, we present an empirical example illustrating the application of the recommended frequentist methods. The empirical example is from a study by Sanchez et al. (2017) for which all materials and data are publicly available through Open Science Framework and can be accessed at (<https://osf.io/g5fvw/>). The code and detailed analysis results for the example is available in the supplemental materials. As mentioned earlier, not all these methods have been thoroughly examined in testing indirect effects in two-mediator sequential mediation chains for both multivariate normal and nonnormal data. The results of the simulation study should help guide best practices for applications of sequential mediation models. We believe that this is one of the largest and most comprehensive Monte Carlo simulation studies evaluating sequential mediation models.

Tests of indirect effects in sequential mediation models

Nonparametric bootstrap

To compute a $(1 - \alpha)100\%$ CI for an indirect effect, denoted by $\theta = \beta_1 \beta_2 \beta_3$, the nonparametric bootstrap draws R repeated samples (i.e., bootstrap samples; $R \geq 1,000$ is recommended, Shrout & Bolger, 2002) with replacement from the original data set (Bollen & Stine, 1990; Efron & Tibshirani, 1993). A mediation model is fitted to the original data to provide an estimate of the indirect effect, where $\hat{\theta} = \hat{\beta}_1 \hat{\beta}_2 \hat{\beta}_3$ denotes the ML estimate of the original sample. The indirect effect is also computed for each bootstrap sample resulting in $\theta_1^*, \theta_2^*, \dots, \theta_R^*$, where θ_r^* denotes the indirect effect estimate for the r th bootstrap sample, to approximate the sampling distribution of the estimated indirect effect and to compute a $(1 - \alpha) 100\%$ CI for the population indirect effect. The percentile method uses $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap samples to obtain the confidence limits.

Efron (1987) also proposed the bias-corrected (BC) bootstrap procedure to account for the median bias (difference between median and mean) of the bootstrap samples. In addition, bias-corrected and accelerated (BCa) bootstrap was proposed to correct for skewness and to yield more accurate coverage for smaller sample sizes (Chernick & LaBudde, 2011; Davison & Hinkley, 1997). Both methods compute adjusted percentiles α'_1 and α'_2 instead of $\alpha/2$ and $1 - \alpha/2$. The adjusted percentiles α'_1 and α'_2 are then used to compute new confidence limits from the bootstrap sample. In both methods, the first step is to calculate the proportion of the bootstrap indirect effect estimates, denoted by p^* , that are less than the original sample estimate $\hat{\theta}$. Then, p^* and $\alpha/2$ quantiles of the standard normal distribution are obtained, and $z_0 = \Phi^{-1}(p^*)$, $z_{\alpha/2} = \Phi^{-1}(\alpha/2)$, and $z_{1-\alpha/2} = -z_{\alpha/2}$, are computed, where Φ^{-1} denotes the inverse of the cumulative standard normal distribution function (e.g., $\Phi^{-1}(.025) = -1.96$). Note that z_0 is an estimate of bias. Next, adjusted percentiles α'_1 and α'_2 are computed for each method. For the BC interval, the adjusted percentiles for the lower and upper CI limit are as follows: $\alpha'_1 = \Phi(2z_0 + z_{\alpha/2})$ and $\alpha'_2 = \Phi(2z_0 + z_{1-\alpha/2})$, where Φ is the cumulative standard normal distribution function. For the BCa interval, the adjusted percentiles are $\alpha'_1 = \Phi(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})})$ and $\alpha'_2 = \Phi(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})})$, where a is an “acceleration” constant that can be estimated during the bootstrapping process or using a jackknife method and adjusts for skewness (Davison & Hinkley, 1997).

Parametric (Monte Carlo) bootstrap

The MC-ML method, also known as the parametric bootstrap (Efron & Tibshirani, 1993), is a flexible method that can be extended to estimate CIs to test sequential mediation models (Tofighi & MacKinnon, 2016). To implement MC-ML, first the posited mediation model is estimated using the ML method. Then, R ($\geq 1,000$) random samples are drawn from a multivariate normal distribution whose mean equals the ML coefficient estimates from the fitted model and its covariance matrix equals the ML estimate covariance matrix of the coefficients. Monte Carlo sample of the indirect effect equals the product of the Monte Carlo sample of the coefficients that comprise the indirect effect. The limits of a $(1 - \alpha)100\%$ CI are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the Monte Carlo sample of the indirect effects.

One variation of the Monte Carlo method that has not been studied for a two-mediator sequential model is MC-Robust (Falk, 2018). We will study the MC-Robust CI with robust Huber-White (Huber, 1967; White, 1980) standard errors (and robust covariance of the parameter estimates) to adjust for the potential nonnormality of data. Falk (2018) studied MC-Robust with robust Satorra and Bentler (2010) standard error correction for single mediator model. The difference between MC-ML and robust MC-Robust is that the latter uses the robust estimates of standard errors (and covariances) of the parameter estimates rather than the default ML standard errors. “Robust standard errors are estimates of standard errors that are supposedly robust against nonnormality.” (Kline, 2016, p. 238). To obtain the robust estimate of the standard errors, the ML estimate of the covariance of the parameter estimates is adjusted using Huber-White sandwich estimator to correct for potential nonnormality. In the sandwich estimator, the “meat” is the correction matrix that is pre- and post-multiplied by the “bread,” which is the ML estimate of the covariance matrix (Huber, 1967; Savalei, 2014; White, 1980). As Freedman (2006) explains, “the sandwich algorithm, under stringent regularity conditions, yields variances for the MLE [maximum likelihood estimators] that are asymptotically correct even when the specification—and hence the likelihood function—are incorrect” (p. 302). Thus, it can be quite a useful way to approximate something that otherwise is unknown.

Model-based (semi-parametric) bootstrap

One potential issue with the nonparametric bootstrap techniques is that the bootstrap samples are drawn from raw data without any consideration for the hypothesized mediation model. Bollen and Stine (1990) proposed a semi-parametric, model-based bootstrap, which is also known as Bollen-Stine (BS) bootstrap. In BS bootstrap, first the sample data is transformed to mimic the population data. Next, bootstrap samples are drawn from the transformed sample data. Then, each sample is used to estimate the model and obtain R bootstrap samples of the quantity of interest. Finally, a CI is obtained by finding the lower and upper quantiles of the bootstrap sample.

To discuss the transformation more specifically, let y_i denote the $p \times 1$ vector of observations for person i , let \mathbf{S} be the sample covariance matrix, let Σ be the hypothesized “population” covariance matrix implied by the mediation model, and let $\hat{\Sigma}$ be the ML estimate

of the hypothesized covariance matrix. The BS bootstrap data is transformed before resampling as follows: $\mathbf{z}_i = \hat{\Sigma}^{1/2} \mathbf{S}^{-1/2} \mathbf{y}_i$; the superscript -1 denotes the inverse matrix and superscript $1/2$ denotes a square root of the positive definite matrix, \mathbf{M} , such that $(\mathbf{M}^{1/2})^T \mathbf{M}^{1/2} = \mathbf{M}$. Note that the transformation is performed to ensure that the covariance matrix of the transformed data equals that of the estimated hypothesized population: $\text{cov}(\mathbf{z}_i) = \hat{\Sigma}$. As a result, the transformation assumes an “exact” fit of data to the hypothesized population.

YHY method (Yuan et al., 2007), an extension of BS bootstrap, was developed to accommodate an approximate fit between the sample and the hypothesized population model. That is, instead of using $\hat{\Sigma}$ to transform data, YHY method uses the following covariance matrix $\mathbf{S}_a = a \mathbf{S} + (1-a) \hat{\Sigma}$, $0 < a < 1$, where a is a constant that is estimated through a numerical algorithm. \mathbf{S}_a can be thought of as a weighted average between the sample covariance matrix and estimated population matrix. Data is transformed as follows: $\mathbf{z}_i = \mathbf{S}_a^{1/2} \mathbf{S}^{-1/2} \mathbf{y}_i$. YHY method resamples the transformed data to achieve an approximate fit between the sample and hypothesized covariance matrix instead of the exact fit between the two.

Profile likelihood

The profile-likelihood approach (Cheung, 2007; Meeker & Escobar, 1995; Pawitan, 2001; Pek & Wu, 2015) produces a CI using the likelihood function, which is the product of the likelihoods for each data point given a specified probability distribution. To compute a profile-likelihood CI, the maximum likelihood estimates are obtained, which is done by maximizing the logarithm of the likelihood function for the model (software does this). Let the vector $\boldsymbol{\theta}$ contain the hypothesized mediation model parameters and $L(\boldsymbol{\theta})$ denote the likelihood function. The log-likelihood function is defined as $LL(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ and the maximum of the log-likelihood function is denoted by LL_1 ,

$$LL_1 = \max LL(\boldsymbol{\theta}) = LL(\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$.

Next, the profile log-likelihood is formed by first assuming that the magnitude of the indirect effect is known, $LL(\boldsymbol{\psi} | \text{IE})$, where $\text{IE} = \beta_1 \beta_2 \beta_3$ stands for indirect effect; and then the profile log-likelihood function is maximized over the unknown “nuisance” parameters denoted by $\boldsymbol{\psi}$:

$$LL_0(\text{IE}) = \max LL(\boldsymbol{\psi} | \text{IE}) = LL(\hat{\boldsymbol{\psi}}_0 | \text{IE}),$$

where the nuisance parameters are the other parameters in the models that are not presented in computing the indirect effect quantities (e.g., β_4 and β_5). The function $LL_0(\text{IE})$, is a profile log-likelihood function that depends on the fixed, but unknown values of IE. Note that the ML estimate $\hat{\boldsymbol{\psi}}_0$ depends on fixed but unknown values of IE. The profile log-likelihood function can be treated as any log-likelihood function. For example, one can compare the profile log-likelihood function to the original log-likelihood function, LL . The parameter space in a profile log-likelihood function is a subset of the original model parameter space because the value of indirect effect was assumed to be fixed. Asymptotically, the following expression has a chi-squared distribution with one degree of freedom (Cox & Hinkley, 2000):

$$-2(LL_0(\text{IE}) - LL_1) \sim \chi^2(1).$$

Finally, the lower and upper bounds for the profile-likelihood $(1 - \alpha)100\%$ CI correspond to the minimum and maximum of the indirect effect that satisfy the following inequality: $-2(LL_0(\text{IE}) - LL_1) \leq \chi^2_{\alpha}(1)$, where $\chi^2_{\alpha}(1)$ denotes upper α critical value of the chi-squared distribution with one degree of freedom.

Bayesian approach

The Bayesian approach yields a credible interval for the product of coefficients using the posterior distribution of the indirect effect (Muthén & Asparouhov, 2012; Yuan & MacKinnon, 2009). Each parameter has a prior distribution, which is the researcher’s belief about the distribution of the parameter before data collection. If there are prior estimates of the coefficients from previous studies (e.g., a meta-analysis), the estimates may be used to form prior distributions (called an informative prior). If there is no information available, one may use a distribution that carries vague or general information about the parameters (called a noninformative, uninformative, flat, or diffuse prior).² A weakly informative (regularizing) prior is used to improve computational stability and inference about a parameter (McElreath, 2016). The Bayesian approach combines model parameter estimates from the current study with the prior distributions to estimate the posterior distribution of all model parameters. A posterior distribution is a

²Although the terms “diffuse” or “uninformative” might be more appropriate in referring to a noninformative prior in our context, we use the term “flat” prior to be consistent with the terminology used in the *rstanarm* package. In our context, a flat prior for a regression coefficient does not mean a uniform prior, but it is a normal distribution with the mean of 0 and standard deviation of 10.

conditional probability distribution that combines the prior distribution with the likelihood from the observed data. Posterior distribution of the parameter is used to compute an interval estimate for each of the parameters. The mechanism of combining the prior distribution and the observed data is known as Bayes' theorem.

When an analytic approach to estimating a posterior distribution of the parameters is not available, the Bayesian approach simulates a random sample from the posterior distribution using Markov Chain Monte Carlo procedures (MCMC; Gilks, Richardson, & Spiegelhalter, 1998; Metropolis & Ulam, 1949). Consider the use of the Bayesian method to estimate the sequential mediation model shown in Figure 1. Random draws (usually in the thousands) from the posterior distribution of all the parameters in the sequential mediation chain are taken. The product of the corresponding coefficients in the indirect effect is computed for each random draw. To create a $(1 - \alpha)100\%$ credible interval, the quantiles that correspond to the lower and upper $\alpha/2$ of the draws from the posterior distribution of the indirect effect are located. Note that MCMC describes a general family of techniques used to draw random samples from a posterior distribution. A few specific MCMC algorithms include Metropolis-Hastings (MH; Hastings, 1970; Metropolis & Ulam, 1949), Gibbs sampling (Geman & Geman, 1984), and Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, & Roweth, 1987; Neal, 2011). For our simulation study, we use the *rstanarm* package (Stan Development Team, 2018) within R, which implements Hamiltonian Monte Carlo algorithm. One advantage of Hamiltonian Monte Carlo compared to MH and Gibbs sampling algorithms is that it requires a fewer number of draws from the posterior distribution (McElreath, 2016).

In the simulation study, we consider two sets of priors for the coefficients that are available in *rstanarm* to estimate the Bayesian 95% credible interval: a weakly informative prior and a flat (noninformative) prior. A weakly informative prior is designed to "provide some information on the relative a priori plausibility of the possible parameter values, for example when we know enough about the variables in our model that we can essentially rule out extreme positive or negative values." (Muth et al., 2018, p. 150). Weakly informative priors have been recommended because they could provide computational stability by regulating the range of the parameter values to prevent extreme values (McElreath, 2016; Stan Development Team, 2018). We assume the priors to

be independent for each coefficient of the indirect effect such that $p(\beta_1, \beta_2, \beta_3) = p(\beta_1) p(\beta_2) p(\beta_3)$, where $p(\beta_1, \beta_2, \beta_3)$ is the joint prior distribution. The weakly informative prior for each coefficient is a normal distribution with the mean of 0 and standard deviation of 2.5, $N(0, 2.5^2)$, which is the default prior in *rstanarm* package. The standard deviation of the prior is automatically adjusted based on the actual range of the dependent variables to ensure that the rescaled prior is weakly informative (Gabry & Goodrich, 2018). We also considered a flat (noninformative) prior for the regression coefficients that assumes a wide range of positive and negative values to be equally likely for the coefficients. In *rstanarm*, the default flat prior for regression coefficients is a normal distribution with the mean of 0 and standard deviation of 10, $N(0, 10^2)$; *rstanarm* rescales the standard deviation based on the actual range of the dependent variables in the model to ensure that the rescaled prior covers a wide range of parameter values (Gabry & Goodrich, 2018). Because of the large variance, the density over the most likely parameter values is approximately flat; this distribution presumably conveys little information about the coefficients and can be thought of as noninformative. For the residual standard deviations of the residual terms (σ_{M_1} , σ_{M_2} , & σ_Y), we used the exponential distribution, denoted by $\exp(\lambda=1)$, where λ is a rate parameter that equals one. Note that this is a default weakly informative prior in *rstanarm*. The parameter λ is also automatically rescaled based on the range of the dependent variables.

Simulation

The purpose of the Monte Carlo simulation study was to empirically assess the Type I error rate, statistical power, and coverage rates of 10 methods of constructing a 95% CI to test the indirect effect ($H_0: \beta_1 \beta_2 \beta_3 = 0$) for the two-mediator sequential model shown in Figure 1. Based on the review of previous simulation studies of mediation models, we manipulated the following four factors in a fully factorial design: (a) distribution, (b) coefficients, (c) sample size, and (d) method of testing (i.e., the 10 tests) the indirect effect. We describe the levels of each factor below.

Distribution

Based on the previous simulation studies of different levels of nonnormality, we considered three multivariate distributions in which we generated multivariate

data to obtain three levels of skewness and kurtosis for each variable (Curran, West, & Finch, 1996; Finch et al., 1997; Nevitt & Hancock, 2001). We first considered a multivariate normal distribution that implies skewness = kurtosis = 0 for each variable, which represents an ideal condition in which all the methods are expected to show their optimal statistical properties. The second condition represents a “moderate” multivariate nonnormal distribution with the marginal univariate skewness of 2 and kurtosis of 7 for each variable. The third condition corresponds to an “extreme” multivariate nonnormal distribution with the univariate skewness of 3 and kurtosis of 21.

Coefficients

The second factor manipulated in the Monte Carlo simulation study was the combination of three coefficients, β_1 , β_2 , and β_3 , which were computed using effect size values of semi-partial R^2 for values of the endogenous variables (mediators and outcome variables). For the two-mediator model in Figure 1, there are three semi-partial R^2 s for the endogenous variables: R^2_{M1} , R^2_{M2} , and R^2_Y . Following Thoemmes et al. (2010), we used these effect sizes to compute the corresponding population values of the β coefficients used to compute the indirect effects (i.e., $\beta_1 \beta_2 \beta_3$). Previous simulation studies (e.g., Biesanz et al., 2010; Taylor et al., 2008) used Cohen’s (1988) guidelines on R^2 effect sizes: 0.02 (small), 0.13 (medium), and 0.23 (large). In addition, our review of the effect sizes of mediation studies in the literature reported semi-partial R^2 larger than 0.23, for example, 0.334 (Adamczyk, 2018) and 0.439 (Huertas-Valdivia, Llorens-Montes, & Ruiz-Moreno, 2018). Thus, we chose the following values for the coefficients: 0, 0.14, 0.36, 0.48, and 0.6. Note that to study the empirical Type I error, one or more of the β coefficients was set to zero while for the power study, none of the coefficients were zero. More specifically, for the Type I error simulation studies, we considered three conditions for the β coefficients: one coefficient equals zero where $\beta_1 \neq 0$, $\beta_2 \neq 0$, and $\beta_3 = 0$; two coefficients equal zero where $\beta_1 \neq 0$ and $\beta_2 = \beta_3 = 0$; and all coefficients equal zero where $\beta_1 = \beta_2 = \beta_3 = 0$. We did not consider other possible combinations such that one coefficient equals zero or two coefficients equal zero because our preliminary simulation studies indicated the results did not depend on the order of the coefficient being zero. For example, for the condition where one coefficient equals zero, preliminary simulation results were virtually the same for the

following conditions: $\beta_1 \neq 0$ and $\beta_2 = \beta_3 = 0$, $\beta_2 \neq 0$ and $\beta_1 = \beta_3 = 0$, and $\beta_3 \neq 0$ and $\beta_1 = \beta_2 = 0$.

Sample size

The third factor manipulated in the Monte Carlo simulation study was sample size (N), which took on the following values: 50, 100, 200, and 500 across all the other manipulated factors. These values of sample size bracket the most commonly used values of sample size seen in empirical studies in psychology and other behavioral sciences. The median, first quartile, and third quartile of the sample sizes from our literature were 209, 110.5, and 359, respectively. A sample size of 50 is generally too small to provide an adequate test of mediation. We used a sample size value of 50 as one of the smaller sizes found in our literature review (e.g., Graham, Martin Ginis, & Bray, 2017). A sample size of 200 roughly equals the median sample size used in the behavioral sciences in studies of regression and SEM (Jaccard & Wan, 1995; MacCallum & Austin, 2000), which is also close to the median sample size of 209 in our literature review. We chose $N = 500$, which was the 85th percentile of the sample sizes in our literature review, as the upper limit because our preliminary simulation study results showed that sample sizes larger than 500 did not provide additional insight about the performance of the methods.

Method

The fourth factor manipulated in the Monte Carlo simulation study, method, was the 10 methods of calculating a 95% CI test of an indirect effect.

Study designs and data generation

The outcomes of simulation study were the Type I error rate, statistical power, and CI coverage. The Type I error rate was measured as the proportion of times the CI does not include zero and hence falsely rejects the null hypothesis of zero indirect effects, when the population value of the indirect effect is in fact zero. Statistical power was measured as the proportion of times a test correctly rejected the null hypothesis of zero indirect effect when the population value of the indirect effect is in fact non-zero. CI coverage is the portion of times the CI included the population value for the indirect effect. For the Type I error rate assessment, at least one of the coefficients must be zero (i.e., no mediation), which resulted in

studying a total of 6,120 conditions. For statistical power, none of the three coefficients are zero, which resulted in 7,680 combinations of non-zero coefficients, distribution, sample size, and method. For the coverage study, we thus considered all 13,800 conditions. We know of no other Monte Carlo simulation study on mediation that examined as many conditions. Thus, the findings we report are the most comprehensive that we are aware of.

Consistent with previous simulation studies of a two-mediator sequential model (e.g., Cheung, 2007), we chose a model in which the relationship between the independent variable on the outcome variable was fully mediated through both M_1 and M_2 ; we assumed the following coefficients to be zero in Figure 1, $\beta_4 = \beta_5 = \beta_6 = 0$. Because we use semi-partial R^2 to generate data, zero versus non-zero values of the coefficients β_4 , β_5 , and β_6 do not change the results in terms of the statistical properties of the tests of indirect effect (Williams & MacKinnon, 2008), which also were supported by our pilot simulation study. We generated data using the population model in Figure 1 based on the combinations of the β coefficients, sample sizes, and skewness and kurtosis values as mentioned earlier. We used *simulateData* function in the *lavaan* (Version 0.6-1) package (Rosseel, 2012) to generate multivariate normal and nonnormal data such that each variable has the specified skewness and kurtosis value (Vale & Maurelli, 1983). The independent variable (X) had a standard normal distribution. The intercepts (not depicted in Figure 1) were all set to zero. For each combination of factors (i.e., Distribution \times $\beta \times N \times$ Method), 1,000 independent replication datasets were generated.

For each replication dataset, we estimated the two-mediator model in Figure 1 in which β_4 , β_5 , and β_6 were constrained to zero. For percentile, BC, BS, and YHY bootstrap, we used *lavaan* built-in functions with 1,000 bootstrap samples as recommended by Shrout and Bolger (2002). For BCa method, we first estimated 1,000 bootstrap samples in *lavaan* and then used the *boot* package (Canty & Ripley, 2017) in R to compute the confidence limits. To calculate MC-ML and MC-Robust CIs, we first estimated the mediation model in *lavaan* and *OpenMx* with regular ML standard error and robust Huber-White standard error, respectively. The parameter estimates and their covariance matrices (ML and Huber-White method) from the estimated mediation models were input into the *ci* function in the *RMediation* (Version 1.1.4) package. We used $R = 100,000$ Monte Carlo samples to

compute the Monte Carlo CIs, which assured a minimum desired accuracy of 0.00001.³ *OpenMx* (Version 2.9.6) was used to compute a profile-likelihood CI for the indirect effects (Neale et al., 2016).

For Bayesian credible intervals, we used two sets of priors for the regression coefficients that are available in *rstanarm* (Stan Development Team, 2018): Bayes-Weak, where the weakly informative prior is $N(0, 2.5^2)$, and Bayes-Flat where the flat prior is $N(0, 10^2)$. Then, as previously noted, *rstanarm* rescales the standard deviations based on the actual range of values of the endogenous variables to ensure that the rescaled prior is weakly informative and noninformative, respectively (Gabry & Goodrich, 2018). For the standard deviations of the residual terms we used the default weakly informative prior, which is the exponential distribution, $\exp(\lambda = 1)$, in *rstanarm*; the rate parameter λ is automatically rescaled based on the range of the endogenous variables to make the prior weakly informative.

Results

Because of the large number of conditions, to save space we only show a subset of the results. More tables and figures are shown in the supplemental materials. Nevertheless, the results we present are indicative of the general set of results.

Type I error rate and accuracy

Multivariate normal distribution

Table 1 shows the Type I error rates for a subset of parameters $\beta_1 = \beta_2$, and $\beta_3 = 0$ (see the supplemental materials for more tables as well as the trellis plots showing trend in the Type I error rate for different sample sizes). To assess the accuracy of the Type I error rates, we use Bradley's (1978) liberal criterion, .025 and .075. The darkest shade of gray shows the inflated Type I error rate above Bradley's upper limit, whereas the lightest shade of gray shows the conservative Type I error rate below Bradley's lower limit. Medium gray shade (between light and dark gray) shows the accurate Type I error rate within the limit. It appears that except for BC and BCa, all the other methods exhibit comparable Type I error rates. BC and BCa showed inconsistent Type I error rates with

³To our knowledge, there is no established guideline for the number of the Monte Carlo samples in mediation analysis. We used *RMediation* to calculate the desired precision of the estimates of the standard errors of the indirect effect. We then conducted preliminary analyses to decide on the number of Monte Carlo samples, conservatively choosing 100,000 Monte Carlo samples to insure stable results.

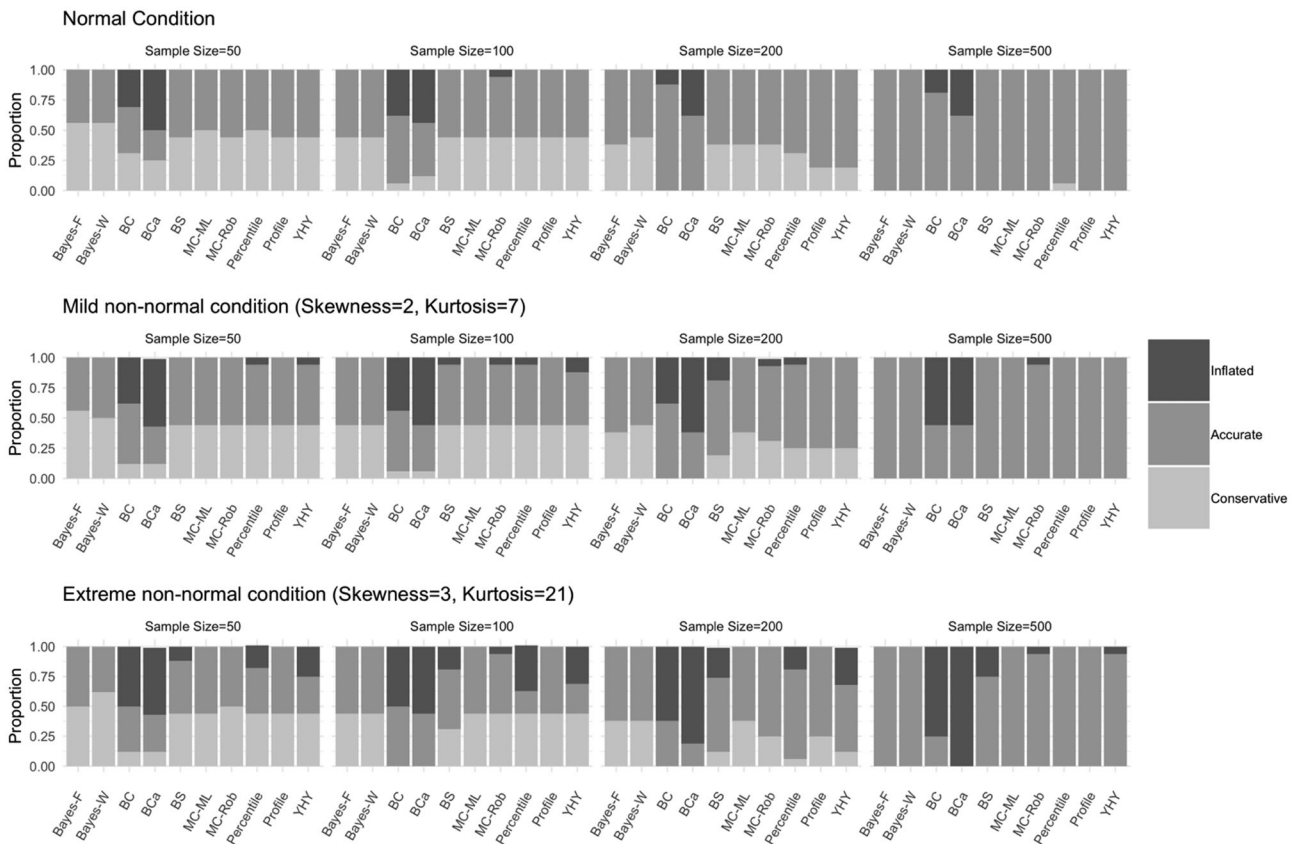


Figure 2. Stacked bar graphs showing accuracy of the Type I Error rate proportions to test an indirect effect, $\beta_1 \beta_2 \beta_3$, where $\beta_3 = 0$. Marginal proportions are calculated for each combination method and sample size averaging across all combinations of non-zero β_s factors ($4 \times 4 \times 1,000$ replications). Using Bradley's (1978) robustness interval for $\alpha = .05$, [.025, .075], marginal proportion of inflated Type I errors is the proportion of the simulation replications for each combination of method and sample size that exceeds 0.075. Marginal proportion of accurate Type I errors is the proportion of the replications that fall within Bradley's interval. Marginal proportion of conservative Type I errors is the proportion of the replications that are less than .025. Bayes-F = Bayesian credible interval with flat (noninformative) prior; Bayes-W = Bayesian credible interval with weakly informative prior; BC = Bias corrected bootstrap; BCa = Bias corrected and accelerated bootstrap; BS = Bollen-Stine semi-parametric bootstrap; MC-ML = Monte Carlo with ML standard errors; MC-Rob = Monte Carlo with robust standard errors; Percentile = Percentile bootstrap; Profile = Profile-likelihood; YHY = Yuan-Hayashi-Yanagihara semi-parametric bootstrap.

multiple instances below and above Bradley's criteria for $N \leq 200$. For $N = 500$, BC and BCa both showed inflated Type I error rates for smaller values of non-zero β_s . All other methods tend to become more accurate as the sample size and magnitude of non-zero β_s increases. The stacked bar-graph in Figure 2 provides the proportion of times a method showed inflated, accurate, and conservative Type I error rates according to Bradley's criterion. Except for BC and BCa, all the other eight methods provide comparable performance in terms of the highest proportion of accurate Type I error rates and the lowest proportion of inflated Type I error rates. As the sample size and magnitude of non-zero coefficients increased, the eight methods tend to become more accurate and less conservative. For $N = 50$, the BC (BCa) bootstrap showed the worst performance with the highest inflation rate of 31% (50%) and lowest accuracy of 38%

(25%) compared to the other methods. For $N = 100, 200, \text{ and } 500$, the BC (BCa) bootstrap method showed inflated Type I error percentages of 38% (44%), 12% (38%), and 19% (38%), respectively. When two of the coefficients were zero with another coefficient non-zero, such as when $\beta_2 = 0, \beta_3 = 0$, and $\beta_1 \neq 0$, all 10 methods were conservative (e.g., empirical Type I error rate around .0025 instead of the nominal Type I error rate of .05), showing the Type I error rate below the lower limit of Bradley's interval.

Multivariate nonnormal distribution

For a subset of conditions, Table 1 also shows the Type I error rate for the moderate multivariate non-normality condition (skewness = 2, kurtosis = 7) and extreme nonnormality condition (skewness = 3, kurtosis = 21), respectively (see the supplemental materials for additional tables and graphs). Compared to the

Type I Error Rates for a Subset of Conditions where $\beta_1 = \beta_2$ and $\beta_3 = 0$.

$\beta_1 = \beta_2$	N	Bayes-F	Bayes-W	BC	BCa	BS	MC-ML	MC-Rob	Percentile	Profile	YHY
Normality Condition											
0.14	50	0.001	0.002	0.008	0.010	0.000	0.002	0.002	0.000	0.001	0.001
	100	0.001	0.003	0.011	0.018	0.001	0.005	0.004	0.003	0.001	0.004
	200	0.005	0.006	0.036	0.054	0.010	0.004	0.008	0.009	0.015	0.007
	500	0.037	0.028	0.095	0.089	0.027	0.038	0.031	0.023	0.039	0.026
0.36	50	0.010	0.021	0.067	0.100	0.025	0.020	0.056	0.023	0.046	0.025
	100	0.031	0.033	0.092	0.116	0.046	0.037	0.046	0.042	0.038	0.057
	200	0.056	0.053	0.061	0.100	0.046	0.057	0.038	0.057	0.053	0.045
	500	0.063	0.065	0.074	0.055	0.045	0.056	0.054	0.050	0.048	0.057
0.48	50	0.038	0.042	0.090	0.126	0.044	0.054	0.067	0.042	0.046	0.047
	100	0.060	0.050	0.071	0.086	0.061	0.067	0.060	0.049	0.054	0.050
	200	0.040	0.053	0.068	0.080	0.052	0.051	0.056	0.056	0.046	0.053
	500	0.060	0.060	0.045	0.061	0.061	0.048	0.052	0.059	0.060	0.060
0.62	50	0.035	0.056	0.086	0.100	0.072	0.058	0.071	0.052	0.062	0.068
	100	0.057	0.039	0.068	0.056	0.050	0.061	0.038	0.054	0.045	0.052
	200	0.058	0.049	0.053	0.052	0.051	0.060	0.052	0.051	0.053	0.057
	500	0.045	0.035	0.052	0.055	0.051	0.053	0.060	0.042	0.048	0.054
Moderate Non-normality Condition (skewness = 2, kurtosis = 7)											
0.14	50	0.001	0.001	0.004	0.009	0.001	0.000	0.000	0.004	0.005	0.002
	100	0.002	0.003	0.024	0.010	0.002	0.007	0.000	0.001	0.006	0.002
	200	0.006	0.004	0.048	0.058	0.009	0.006	0.006	0.008	0.016	0.006
	500	0.034	0.025	0.107	0.097	0.034	0.026	0.031	0.033	0.026	0.054
0.36	50	0.011	0.016	0.075	0.096	0.030	0.026	0.029	0.038	0.034	0.043
	100	0.038	0.039	0.114	0.104	0.065	0.031	0.046	0.059	0.055	0.057
	200	0.062	0.049	0.074	0.096	0.063	0.052	0.058	0.061	0.050	0.072
	500	0.057	0.049	0.069	0.075	0.051	0.039	0.075	0.055	0.056	0.061
0.48	50	0.036	0.033	0.085	0.127	0.070	0.041	0.058	0.071	0.049	0.069
	100	0.043	0.040	0.094	0.116	0.077	0.067	0.079	0.058	0.061	0.066
	200	0.035	0.048	0.076	0.079	0.077	0.061	0.060	0.064	0.046	0.068
	500	0.049	0.052	0.066	0.096	0.053	0.048	0.071	0.069	0.050	0.048
0.62	50	0.053	0.045	0.072	0.091	0.073	0.047	0.063	0.078	0.055	0.079
	100	0.050	0.050	0.071	0.089	0.075	0.040	0.071	0.061	0.054	0.073
	200	0.046	0.049	0.070	0.086	0.065	0.044	0.081	0.067	0.045	0.060
	500	0.066	0.051	0.058	0.072	0.066	0.061	0.048	0.056	0.053	0.065
Extreme Non-normality Condition (skewness = 3, kurtosis = 21)											
0.14	50	0.000	0.003	0.010	0.014	0.002	0.002	0.000	0.001	0.005	0.001
	100	0.003	0.005	0.026	0.027	0.007	0.002	0.000	0.005	0.006	0.003
	200	0.007	0.008	0.048	0.085	0.010	0.010	0.008	0.009	0.014	0.021
	500	0.030	0.027	0.108	0.125	0.047	0.027	0.033	0.046	0.048	0.054
0.36	50	0.023	0.022	0.078	0.093	0.045	0.034	0.019	0.051	0.034	0.061
	100	0.036	0.053	0.116	0.132	0.069	0.045	0.048	0.059	0.056	0.073
	200	0.067	0.058	0.096	0.110	0.068	0.046	0.063	0.071	0.053	0.085
	500	0.058	0.051	0.072	0.097	0.055	0.052	0.060	0.074	0.047	0.067
0.48	50	0.042	0.046	0.104	0.124	0.084	0.043	0.073	0.073	0.056	0.079
	100	0.058	0.046	0.082	0.120	0.072	0.050	0.063	0.080	0.044	0.088
	200	0.047	0.046	0.096	0.117	0.102	0.058	0.058	0.071	0.064	0.070
	500	0.037	0.051	0.080	0.086	0.047	0.047	0.075	0.068	0.049	0.067
0.62	50	0.061	0.050	0.092	0.102	0.071	0.060	0.071	0.077	0.060	0.077
	100	0.056	0.053	0.073	0.114	0.085	0.062	0.056	0.085	0.043	0.090
	200	0.043	0.055	0.087	0.086	0.076	0.040	0.058	0.061	0.057	0.090
	500	0.053	0.063	0.079	0.088	0.069	0.034	0.048	0.070	0.054	0.073

Note. Darkest shade of gray shows the inflated Type I error rate above Bradley's (1978) upper limit (.075) while the lightest shade of gray shows the conservative Type I error rate below Bradley's lower limit (.025). Medium gray (between light and dark gray) shows the accurate Type I error rate within the limit, [.025, .075].

Bayes-F = Bayesian credible interval with flat (noninformative) prior; Bayes-W = Bayesian credible interval with weakly informative prior; BC = Bias corrected bootstrap; BCa = Bias corrected and accelerated bootstrap; BS = Bollen-Stine semi-parametric bootstrap; MC-ML = Monte Carlo with ML standard errors; MC-Rob = Monte Carlo with robust standard errors; Percentile = Percentile bootstrap; Profile = Profile-likelihood; YHY = Yuan-Hayashi-Yanagihara semi-parametric bootstrap.

normal condition results, BC and BCa showed more instances of inflated Type I error rates for the moderate multivariate normality condition, which got worse for the extreme normality condition. For the moderate multivariate nonnormality condition, MC-Robust, percentile, BS, and YHY methods showed multiple instances of inflated Type I error rates; the frequency of the inflated Type I error rate increased for the extreme nonnormality condition. The rest of the

methods provided comparable Type I error rates across the conditions. As the sample size and magnitude of non-zero coefficients increased, all methods except for BC and BCa tend to become more accurate. Similar to the multivariate normal condition results, when two of the coefficients were zero, $\beta_2 = \beta_3 = 0$ and $\beta_1 \neq 0$, all 10 methods were conservative, showing a Type I error rate below the lower limit of Bradley's (1978) interval.

Table 2. Power to detect indirect effect for a subset of conditions, where $\beta_1 = \beta_2 = \beta_3$.

Coefficients		Method to compute 95% confidence/credible interval									
$\beta_1 = \beta_2 = \beta_3$	N	Bayes-F	Bayes-W	BC	BCa	BS	MC-ML	MC-Rob	Percentile	Profile	YHY
Normality Condition											
0.14	50	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.01	0.00
	100	0.01	0.01	0.08	0.09	0.01	0.01	0.02	0.02	0.03	0.02
	200	0.08	0.08	0.25	0.28	0.10	0.11	0.09	0.10	0.14	0.11
	500	0.64	0.64	0.81	0.83	0.67	0.68	0.66	0.65	0.70	0.69
0.36	50	0.37	0.37	0.56	0.62	0.37	0.41	0.41	0.38	0.46	0.39
	100	0.87	0.90	0.94	0.93	0.87	0.90	0.90	0.88	0.91	0.88
	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.48	50	0.85	0.86	0.90	0.93	0.86	0.86	0.87	0.88	0.88	0.86
	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.62	50	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99
	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Moderate Non-normality Condition (skewness=2, kurtosis=7)											
0.14	50	0.01	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.01	0.00
	100	0.02	0.02	0.08	0.08	0.01	0.02	0.01	0.01	0.03	0.02
	200	0.09	0.10	0.27	0.28	0.09	0.11	0.06	0.09	0.16	0.09
	500	0.62	0.60	0.81	0.83	0.66	0.61	0.57	0.67	0.67	0.63
0.36	50	0.33	0.36	0.57	0.57	0.36	0.36	0.28	0.37	0.42	0.39
	100	0.82	0.85	0.92	0.94	0.88	0.82	0.76	0.88	0.87	0.88
	200	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.48	50	0.80	0.79	0.91	0.91	0.85	0.82	0.75	0.85	0.83	0.85
	100	0.99	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.62	50	0.99	0.99	1.00	1.00	1.00	0.99	0.97	0.99	0.99	0.99
	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Extreme Non-normality Condition (skewness=3, kurtosis=21)											
0.14	50	0.00	0.00	0.03	0.03	0.01	0.01	0.01	0.00	0.01	0.01
	100	0.02	0.02	0.09	0.11	0.02	0.01	0.00	0.03	0.04	0.03
	200	0.10	0.10	0.32	0.34	0.13	0.11	0.05	0.15	0.13	0.12
	500	0.60	0.60	0.87	0.86	0.71	0.59	0.56	0.71	0.64	0.74
0.36	50	0.32	0.32	0.59	0.64	0.51	0.33	0.28	0.48	0.43	0.49
	100	0.78	0.79	0.95	0.94	0.92	0.83	0.61	0.93	0.82	0.93
	200	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.48	50	0.76	0.77	0.91	0.91	0.90	0.80	0.60	0.90	0.81	0.91
	100	0.99	0.99	1.00	1.00	1.00	0.99	0.92	1.00	0.99	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.62	50	0.98	0.98	0.99	1.00	1.00	0.99	0.88	1.00	0.99	1.00
	100	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. Bayes-F = Bayesian credible interval with flat (noninformative) prior; Bayes-W = Bayesian credible interval with weakly informative prior; BC = Bias corrected bootstrap; BCa = Bias corrected and accelerated bootstrap; BS = Bollen-Stine semi-parametric bootstrap; MC-ML = Monte Carlo with ML standard errors; MC-Rob = Monte Carlo with robust standard errors; Percentile = Percentile bootstrap; Profile = Profile-likelihood; YHY = Yuan-Hayashi-Yanagihara semi-parametric bootstrap.

Statistical power

We present power results for a subset of conditions where $\beta_1 = \beta_2 = \beta_3$ in Table 2 with horizontal data bars. Data bars are similar to a bar chart in that each bar represents a relative height equal to the power

value in a cell (see the supplemental materials for more results). Data bars make it easier to compare ranges of values. We do not discuss BC and BCa bootstrap further in the power study because they did not meet the necessary condition of showing an

accurate Type I error rate when the null hypothesis was true, thus violating the principles of statistical hypothesis testing (Davison, 2003; Lehmann & Romano, 2005).

Multivariate normal distribution

For $N \geq 200$ and $\beta_s \geq .36$, the power for all eight methods exceeded .96. In addition, as the sample size and size of the indirect effect increased, power for all methods appeared to increase or to remain the same (note that there is a ceiling effect for power). Also, for larger sample sizes and effect sizes, difference in power between the methods tended to decrease. For the small effect size ($\beta_1 = \beta_2 = \beta_3 = 0.14$), one needs at least 500 observations to achieve a minimum power of .64 with either of the Bayesian methods and a maximum power of .70 with the profile-likelihood method; maximum power for $N=200$ was .14 with the profile-likelihood method. The difference between the eight methods did not exceed .09. The maximum difference of .09 occurred for the medium effect size ($\beta_1 = \beta_2 = \beta_3 = 0.36$) and $N=50$ with the maximum of .46 for the profile-likelihood method and the minimum of .37 for the two Bayesian methods and BS bootstrap.

Multivariate nonnormal distribution

For the moderate multivariate nonnormality condition (skewness = 2 and kurtosis = 7), the largest difference in power between the eight methods was .14, which occurred for medium effect size ($\beta_1 = \beta_2 = \beta_3 = 0.36$) and $N=50$; for this condition, the profile-likelihood method had a maximum power of .42 and the MC-ML method had a minimum power of .28. The second largest power difference was .13, which occurred for $N=100$ and the medium effect sizes; for this condition, the MC-ML method had a minimum power of .76 while the percentile bootstrap, BS, and YHY methods had a maximum power of .88. For the extreme nonnormality condition (skewness = 3 and kurtosis = 21), the difference in power between the methods increased compared to moderate multivariate nonnormality and normality conditions. The largest power difference was .33, which occurred for medium effect size and $N=100$; the MC-ML had a minimum power of .61 and the percentile and YHY methods had a maximum power of .93.

Coverage

Table 3 shows coverage values for the 10 methods for a subset of values for β -coefficients ($\beta_1 = \beta_2 = \beta_3$)

and sample sizes for multivariate normality, moderate multivariate nonnormality, and extreme multivariate nonnormality conditions, respectively. To facilitate interpretation, we use Bradley's (1978) criterion to identify under-coverage (<.925), over-coverage (>.975), and accurate coverage (between .925 and .975). Cells with an upward pointing arrow indicate over-coverage while a downward pointing arrow indicates under-coverage; cells with no arrows indicate accurate coverage. Also, Figures 3–5 show jittered dot plots of coverage values for all 10 methods as a function of sample sizes, which are collapsed (averaged) across coefficients values, for multivariate normality, moderate multivariate nonnormality, and extreme multivariate nonnormality conditions, respectively. The dot plots show the distribution of the coverage values and facilitate understanding of how coverage values are distributed in relation to the Bradley's interval, whose limits are drawn with solid lines in each plot. In discussing coverage, under-coverage is considered the less ideal outcome compared to over-coverage (Falk & Biesanz, 2014). Thus, ideally, confidence/credible interval must show empirical coverage that equals or exceeds the nominal value of $1 - \alpha$ (.95) while exhibiting low frequency of under-coverage. Also, there are different degrees of under-coverage in terms of how far coverage falls below the lower limit of Bradley's (1978) criterion.

Multivariate normal distribution

As shown in Table 3 and Figure 3, when $N=50$, the profile likelihood and both Bayesian methods showed the best coverage in terms of accuracy and no under-coverage while BC and BCa frequently showed under-coverage with coverage values of .90 and .89, respectively. For $N=100$, the profile-likelihood, MC-ML, and both Bayesian methods showed the best coverage followed closely by the percentile, BS, and YHY bootstrap. BC and BCa showed under-coverage, with the lowest coverage of .89 and .87, respectively; however, compared to $N=50$, the occurrence of under-coverage was less frequent. For $N=200$, all 10 methods exhibited comparably accurate coverage. Overall, for normal condition, the profile-likelihood and Bayesian methods showed the best coverage followed by the MC-ML, percentile, BS, and YHY methods. We do not recommend BC and BCa, especially for $N \leq 100$.

Multivariate nonnormal distribution

For both multivariate nonnormality conditions (Table 3 and Figures 4 and 5), we arranged the methods into

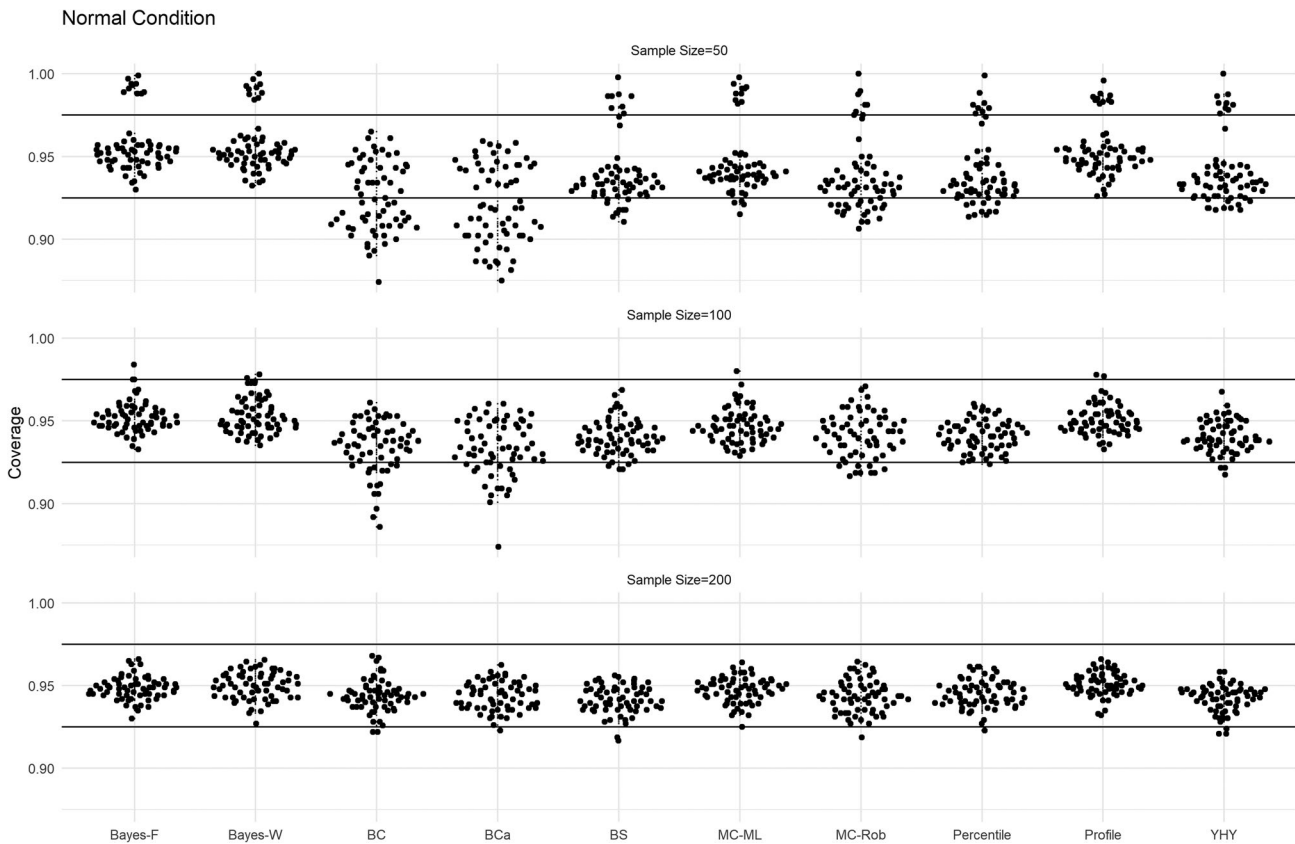


Figure 3. Jittered dotplot of coverage for multivariate normal condition. Bayes-F = Bayesian credible interval with flat (non-informative) prior; Bayes-W = Bayesian credible interval with weakly informative prior; BC = Bias corrected bootstrap; BCa = Bias corrected and accelerated bootstrap; BS = Bollen-Stine semi-parametric bootstrap; MC-ML = Monte Carlo with ML standard errors; MC-Rob = Monte Carlo with robust standard errors; Percentile = Percentile bootstrap; Profile = Profile-likelihood; YHY = Yuan-Hayashi-Yanagihara semi-parametric bootstrap. Solid horizontal lines show the limits of Bradley's (1978) interval, [.925, .975].

three groups sorted from best to worst coverage: (a) percentile bootstrap, YHY, and BS method; (b) BC, BCa, and MC-Robust; (c) profile-likelihood, MC-ML and two Bayesian methods. For both multivariate nonnormality conditions, the methods in the first group showed the best coverage although coverage was worse compared to the normality condition for $N=50$ and 100. These three methods showed comparable coverage within and across the multivariate nonnormality conditions. The minimum coverage for the percentile, YHY, and BC method ranged from .89 to .90 for both multivariate nonnormality conditions. In the second-best group, for moderate multivariate nonnormality condition, all three methods showed comparable coverage that tended to improve as sample size, size of the indirect effect, or both increased. The minimum coverage for the methods ranged from .85 to .86. For the extreme multivariate nonnormality condition, the coverage for BC and BCa remained the same. However, coverage of the MC-Robust degraded by 3% on average. In the third group, all four methods showed the lowest coverage. For the moderate

multivariate nonnormality condition, the minimum coverage ranged from .81 to .82; for the extreme multivariate nonnormality condition, the minimum coverage ranged from .67 to .68. Poorer performance could be because these four methods rely more heavily on multivariate normality distribution of the coefficient estimates without any adjustment for nonnormality. In addition, an interesting result was that coverage for these methods got worse for the larger sizes of indirect effect, sample size, or both; one possible explanation for the poor coverage could be as the sample size increased, the standard errors decreased, and CIs became narrower, thus coverage worsened.

Empirical example

The empirical example is part of a study by Sanchez et al. (2017), for which all of the data and study materials are available to the public via the Open Science Framework (please follow the link <https://osf.io/g5fvw/> to access the materials). Sanchez et al. conducted five

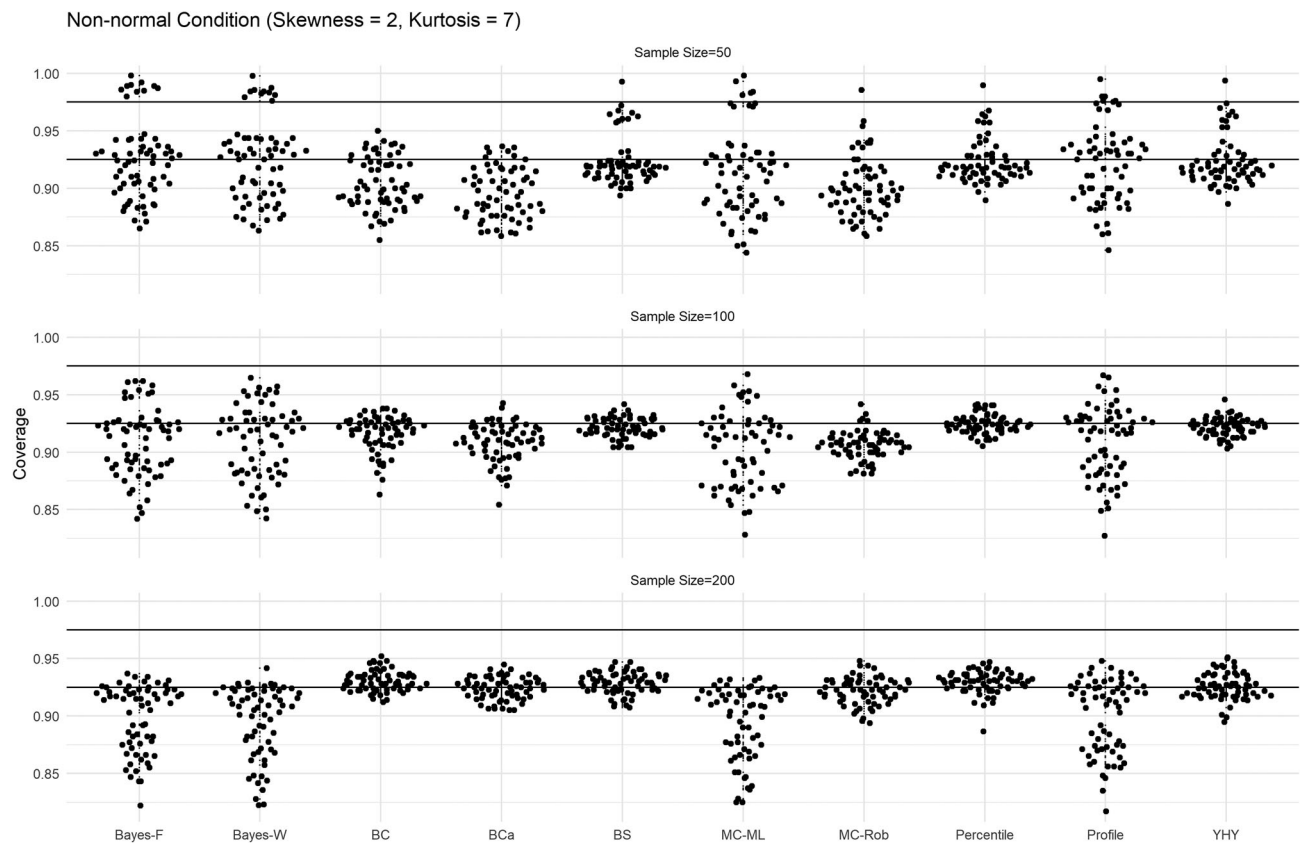


Figure 4. Jittered dotplot of coverage for the moderate multivariate nonnormality condition (skewness = 2, kurtosis = 7). Bayes-F = Bayesian credible interval with flat (noninformative) prior; Bayes-W = Bayesian credible interval with weakly informative prior; BC = Bias corrected bootstrap; BCa = Bias corrected and accelerated bootstrap; BS = Bollen-Stine semi-parametric bootstrap; MC-ML = Monte Carlo with ML standard errors; MC-Rob = Monte Carlo with robust standard errors; Percentile = Percentile bootstrap; Profile = Profile-likelihood; YHY = Yuan-Hayashi-Yanagihara semi-parametric bootstrap. Solid horizontal lines show the limits of Bradley's (1978) interval, [.925, .975].

studies of stigma by transfer with different stigmatized group (e.g., White women). Stigma by transfer means that a member of the stigmatized group is more likely to view racists as sexists and vice versa. We focus on one of the mediation models in Study 1 (there are five studies), in which the authors examined indirect effects of viewing the profiles of racist individuals (Treatment = Racist profile vs control) on gender identity threat (Gender Stigma) via the two sequential mediators (a) perceived social dominance orientation (Perceived SDO) and (b) perceived sexism (Perceived Sexism), as shown in Figure 6.

The data from Study 1 consists of a subset of participants, where Study 1 uses only the Females participants ($N = 100$). The female participants were randomly assigned to view responses to the Modern Racism Scale and the Old Fashioned Racism Scale (McConahay, 1986) from an individual with evidence of a "moderate" racist attitude (Racism condition) and to the neutral profile (e.g., with responses to personality measures) with no evidence of a sexist or a racist attitude (Control condition). To measure Perceived

Sexism, participants responded to question on a 5-item scale (e.g., "How likely is it that this person treats women fairly?") to evaluate the profiled person, in which 1 indicated "very slightly or not at all" and 5 indicated "extremely or a lot" ($\alpha = .967$).

To measure Perceived SDO, the participants completed a 16-item SDO scale (Pratto, Sidanius, Stallworth, & Malle, 1994) as the profiled person would have done ($\alpha = .979$). Response to each item (e.g., "Some groups of people are simply inferior to other groups.") ranged from 1 (very negative or strongly disagree) to 7 (very positive or strongly agree). To measure Gender Stigma ($\alpha = .974$), respondents answered the question about the profile person, "How much would you be concerned that this person would judge you based on the following characteristics?", where the characteristics were "My gender", "My sex", and "My being a woman". The answers ranged from 1 indicating "not at all" to 7 indicating "a great deal". Finally, to measure Liking, a 3-item scale was used, with an example question, "If you were in a room with this person, would you have

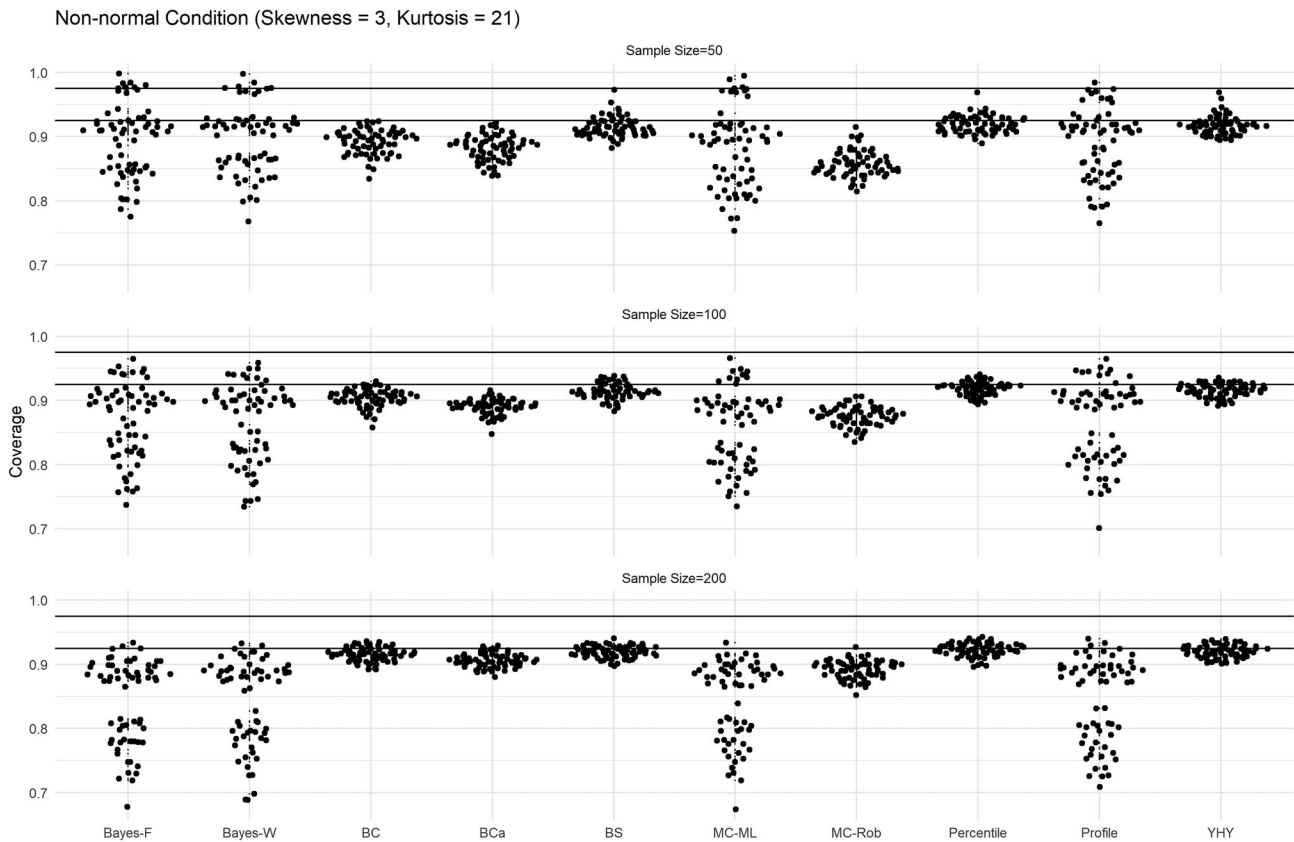


Figure 5. Jittered dotplot of coverage for the extreme multivariate nonnormality condition (skewness, 3, kurtosis, 21). Bayes-F = Bayesian credible interval with flat (noninformative) prior; Bayes-W = Bayesian credible interval with weakly informative prior; BC = Bias corrected bootstrap; BCa = Bias corrected and accelerated bootstrap; BS = Bollen-Stine semi-parametric bootstrap; MC-ML = Monte Carlo with ML standard errors; MC-Rob = Monte Carlo with robust standard errors; Percentile = Percentile bootstrap; Profile = Profile-likelihood; YHY = Yuan-Hayashi-Yanagihara semi-parametric bootstrap. Solid horizontal lines show the limits of Bradley's (1978) interval, [.925, .975].

a lot of things to talk about?" ($\alpha = .778$). The answers ranged from 1 indicating "very slightly or not at all" to 5 indicating "extremely or a lot". For Perceived SDO, Perceived Sexism, and Liking, composite scores of the respected scales were used in the final analysis.

We fit three ordinary least-squares (OLS) regression equations corresponding to the dependent variables (two mediators and one outcome variable) in Figure 6. The OLS regression allows us to compute case residuals,⁴ which we examined using plots (e.g., qq plot) as well as testing for multivariate normality using Henze and Zirkler (HZ; 1990) method, which has been recommended in the literature (Mecklin & Mundfrom, 2005). We also check for univariate normality of the residuals because it is a necessary condition for multivariate normality. Case residuals can

also be used to check for outliers using t -test with Bonferroni adjusted p -value (Cohen, Cohen, West, & Aiken, 2003; Fox, 2016). Skewness (kurtosis) for residuals associated with SDO, Perceived Sexism, and Gender Stigma was -0.9 (1.7), 0.2 (0.2), and -1.1 (3.9), respectively. Mardia's (1970) multivariate measures of skewness and kurtosis for the residuals were 2.57 and 21, respectively. The result of the HZ test, statistic = 1.307, $p = .001$ indicates that we reject the hypothesis of multivariate normality because the p -value is very small. We flagged two observations as outliers using t -test of the studentized residuals for observation 35, Bonferroni $p < .001$, and observation 19, Bonferroni $p = .036$. We removed the two observations from the data and refit the regression equations.⁵ Skewness (kurtosis) for the new residuals associated with SDO, Perceived Sexism, and Gender Stigma was -0.7 (1.4), 0.3 (0.2), and 0.02 (-0.3),

⁴To our knowledge, software packages such as *OpenMx* and *lavaan* do not have built-in functions to produce case residuals. Instead, these packages compute a variety of the residuals that are a function of the difference between the sample and model implied covariance between the dependent (endogenous) variables in the model.

⁵Generally, we do not recommend removing outliers when robust estimators that down weight the outliers are available. To date, *OpenMx* and *lavaan* do not have an estimator robust to outliers.

Table 3. Coverage of 95% intervals for a subset of conditions, where $\beta_1 = \beta_2 = \beta_3$.

Coefficients		Method to compute 95% confidence/credible interval									
$\beta_1 = \beta_2 = \beta_3$	N	Bayes-F	Bayes-W	BC	BCa	BS	MC-ML	MC-Rob	Percentile	Profile	YHY
Normality condition											
0.14	50	↑0.999	↑1.000	0.927	↓0.920	↑0.998	↑0.998	↑1.000	↑0.999	↑0.996	↑1.000
	100	↑0.984	↑0.976	↓0.892	↓0.874	0.966	↑0.980	0.963	0.960	↑0.978	0.968
	200	0.930	0.941	0.937	↓0.923	↓0.917	0.934	0.935	0.934	0.956	↓0.921
	500	0.929	0.940	0.959	↑0.976	0.942	0.929	0.940	0.931	0.939	0.943
	50	0.950	0.935	0.956	0.953	↓0.916	↓0.915	↓0.921	↓0.915	0.949	0.934
0.36	100	0.933	0.959	0.953	0.950	↓0.923	0.954	0.935	0.947	0.947	0.940
	200	0.937	0.941	0.940	0.956	0.951	0.948	0.952	0.945	0.935	0.938
	500	0.947	0.934	0.951	0.950	0.951	0.949	0.952	0.969	0.956	0.946
	50	0.956	0.947	0.954	0.957	0.927	0.942	0.938	0.925	0.959	0.932
	100	0.954	0.958	0.928	0.950	0.939	0.959	0.929	0.934	0.949	0.933
0.48	200	0.944	0.961	0.926	0.940	0.938	0.950	0.933	0.943	0.951	0.935
	500	0.954	0.951	0.949	0.948	0.936	0.946	0.940	0.941	0.953	0.943
	50	0.957	0.951	0.929	0.952	0.933	0.945	↓0.921	0.935	0.948	0.933
	100	0.949	0.950	0.932	0.957	↓0.921	0.951	↓0.923	0.930	0.946	0.946
	200	0.949	0.943	0.936	0.940	0.942	0.954	0.935	0.940	0.954	0.946
500	0.948	0.945	0.948	0.956	0.946	0.951	0.956	0.955	0.954	0.942	
Moderate nonnormality condition (skewness = 2, kurtosis = 7)											
0.14	50	↑0.998	↑0.998	↓0.888	↓0.903	↑0.993	↑0.998	↑0.985	↑0.990	↑0.995	↑0.994
	100	0.962	0.965	↓0.863	↓0.854	0.929	0.968	0.929	0.933	0.967	↓0.920
	200	↓0.915	↓0.922	0.927	↓0.906	↓0.909	↓0.918	↓0.921	↓0.886	0.942	↓0.899
	500	↓0.919	0.931	0.955	0.942	↓0.922	↓0.924	↓0.896	0.935	0.934	↓0.913
	50	↓0.915	↓0.904	0.934	0.932	↓0.920	↓0.897	↓0.894	↓0.914	0.926	↓0.900
0.36	100	↓0.912	↓0.914	0.927	0.929	0.928	↓0.888	↓0.908	0.929	↓0.901	0.926
	200	↓0.892	↓0.886	0.927	↓0.924	0.927	↓0.890	0.944	0.930	↓0.912	↓0.921
	500	↓0.887	↓0.863	0.942	0.950	0.948	↓0.880	0.931	0.939	↓0.896	0.930
	50	↓0.883	↓0.893	↓0.923	↓0.922	↓0.920	↓0.878	↓0.879	0.928	↓0.895	↓0.923
	100	↓0.879	↓0.890	0.938	↓0.920	↓0.924	↓0.880	↓0.896	0.926	↓0.869	↓0.918
0.48	200	↓0.867	↓0.842	0.935	↓0.921	↓0.922	↓0.837	↓0.923	0.933	↓0.856	0.936
	500	↓0.850	↓0.854	0.934	0.930	0.930	↓0.861	0.935	0.948	↓0.851	0.939
	50	↓0.865	↓0.863	↓0.892	↓0.906	↓0.909	↓0.844	↓0.865	↓0.911	↓0.846	0.925
	100	↓0.842	↓0.842	↓0.902	↓0.922	↓0.923	↓0.828	↓0.917	↓0.922	↓0.827	↓0.920
	200	↓0.822	↓0.823	↓0.923	↓0.905	↓0.923	↓0.825	0.931	↓0.923	↓0.817	0.935
500	↓0.813	↓0.812	0.938	0.926	0.930	↓0.823	0.933	0.936	↓0.836	0.944	
Extreme nonnormality condition (skewness = 3, kurtosis = 21)											
0.14	50	↑0.998	↑0.998	↓0.853	↓0.839	0.973	↑0.995	↓0.915	0.969	↑0.984	0.969
	100	0.965	0.959	↓0.871	↓0.848	↓0.894	0.966	↓0.869	↓0.900	0.965	↓0.903
	200	↓0.905	↓0.920	0.928	↓0.923	↓0.900	0.934	↓0.883	↓0.896	0.940	↓0.906
	500	↓0.917	↓0.902	0.954	0.928	↓0.921	↓0.887	↓0.919	0.926	↓0.915	0.926
	50	↓0.896	↓0.869	↓0.924	↓0.906	↓0.907	↓0.868	↓0.854	↓0.908	↓0.883	↓0.919
0.36	100	↓0.864	↓0.825	↓0.906	↓0.899	↓0.915	↓0.862	↓0.865	↓0.915	↓0.846	0.929
	200	↓0.811	↓0.827	↓0.908	↓0.904	0.925	↓0.811	↓0.915	0.927	↓0.831	↓0.911
	500	↓0.783	↓0.809	0.932	↓0.910	0.929	↓0.795	↓0.923	0.939	↓0.775	0.941
	50	↓0.845	↓0.837	↓0.895	↓0.909	↓0.904	↓0.816	↓0.852	0.935	↓0.839	↓0.920
	100	↓0.785	↓0.791	↓0.921	↓0.894	↓0.905	↓0.804	↓0.879	0.941	↓0.777	↓0.923
0.48	200	↓0.778	↓0.755	0.931	↓0.913	↓0.917	↓0.753	↓0.869	↓0.921	↓0.762	↓0.921
	500	↓0.746	↓0.742	0.932	↓0.904	0.932	↓0.731	0.938	0.940	↓0.737	0.931
	50	↓0.775	↓0.801	↓0.897	↓0.893	↓0.905	↓0.753	↓0.833	↓0.908	↓0.765	↓0.923
	100	↓0.737	↓0.734	↓0.900	↓0.902	↓0.906	↓0.735	↓0.900	↓0.922	↓0.701	0.928
	200	↓0.678	↓0.689	↓0.892	↓0.888	↓0.916	↓0.674	↓0.906	↓0.924	↓0.709	0.931
500	↓0.674	↓0.679	↓0.924	↓0.916	0.932	↓0.675	0.929	0.929	↓0.681	0.936	

Note. Upward arrow and italic font show over-coverage (> .975); downward arrow and bold font show under-coverage (< .925). Bayes-F = Bayesian credible interval with flat (noninformative) prior; Bayes-W = Bayesian credible interval with weakly informative prior; BC = Bias corrected bootstrap; BCa = Bias corrected and accelerated bootstrap; BS = Bollen-Stine semi-parametric bootstrap; MC-ML = Monte Carlo with ML standard errors; MC-Rob = Monte Carlo with robust standard errors; Percentile = Percentile bootstrap; Profile = Profile-likelihood; YHY = Yuan-Hayashi-Yanagihara semi-parametric bootstrap.

respectively. As can be seen, removing the outliers helped reduce (the absolute value of) skewness and kurtosis of the residuals for Gender Stigma. Outliers can be one reason for nonnormality. The result of the HZ test, a test statistic = 1.042, $p=0.03$, indicated we reject the hypothesis of multivariate normality. Examining the skewness and kurtosis values, as well as looking at the qq plots of residuals along with p -value = .03, it is unclear if the evidence

against violation of the multivariate normality is as strong as before when the two cases were included. As a result, in conducting mediation analysis, we would consider two scenarios: one scenario where the multivariate normality seems reasonable (i.e., where we did not reject that null hypothesis of normality) and one scenario where the multivariate normality is violated (i.e., where we reject the null hypothesis of normality).

Table 4. Estimates for the two-mediator sequential model of the empirical example ($N = 100$).

Variables		95% CI						
Dependent	Predictor	Estimate	SE	z	p	LL	UL	Semi-partial R^2
Perceived SDO	Treatment	1.51 (b_1)	0.27	5.50	<.001	0.97	2.04	0.162
	Liking	-0.71	0.15	-4.88	<.001	-1.00	-0.43	0.128
Perceived sexism	Perceived SDO	0.26 (b_2)	0.06	4.64	<.001	0.15	0.37	0.086
	Treatment	0.11 (b_4)	0.17	0.66	0.51	-0.23	0.45	0.002
	Liking	-0.52	0.09	-5.80	<.001	-0.69	-0.34	0.135
Gender stigma	Perceived Sexism	1.31 (b_3)	0.18	7.48	<.001	0.97	1.65	0.194
	Perceived SDO	0.05 (b_5)	0.11	0.51	0.61	-0.16	0.26	0.001
	Treatment	0.18 (b_6)	0.30	0.59	0.56	-0.41	0.77	0.001
	Liking	-0.17	0.18	-0.94	0.35	-0.52	0.18	0.003

Note. LL = lower limit; SDO = social dominance orientation; Treatment = 1 (racism), 0 (control); UL = upper limit.

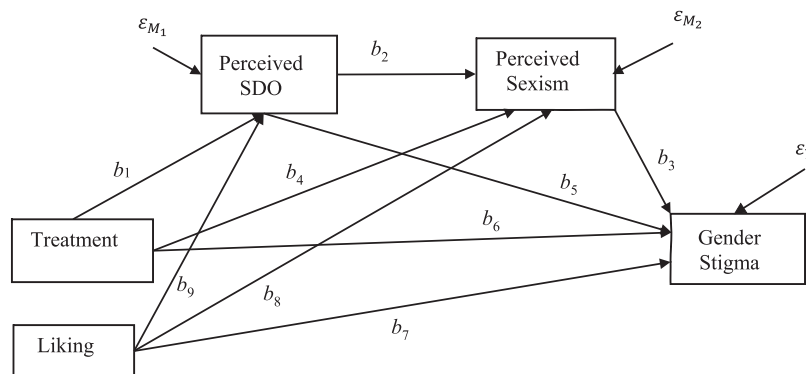


Figure 6. A sequential two-mediator model. The independent variable, Treatment, denotes a random assignment that takes on 1 (Racism), or 0 (Control). The two sequential mediators are perceived social dominance orientation (SDO) and Perceived Sexism. The dependent (outcome) variable is Gender Stigma. The variable Liking is a covariate. The quantity of interest is the indirect effect of Treatment on Gender Stigma that sequentially transmits through both Perceived SDO and Perceived Sexism controlling for the effect of Liking. Under the no-omitted-confounder assumption, the specific indirect effect of Treatment on Gender Stigma through Perceived SDO and Perceived Sexism equals the product of three coefficients, $b_1b_2b_3$.

To compute the 95% CIs for the model in Figure 6, we use *lavaan* and *OpenMx*, each of which have built-in functions to estimate bootstrap and profile-likelihood for an indirect effect, respectively. The results are shown in Table 4. To conduct mediation analysis, our recommendation is based on the purpose (significance testing, reporting an interval estimate, and reporting a model fit) of the mediation analysis and the assumption about the distribution of data (see the Discussion section for additional detail). If one were to assume that the assumption of multivariate normality is not violated, we would use the profile-likelihood, MC-ML, Bayes-Flat or Bayes-Weak methods to conduct both significance testing and to compute CI. We compute the profile likelihood 95% CI: [0.25 0.88] and MC-ML 95% CI: [0.24, 0.87]). Based on the CIs, it appears that indirect effect is different from zero at $\alpha = .05$. Also, it appears that the indirect effect ranges from 0.24 to 0.87 using the MC-ML CI. However, if a researcher were to err on the side of caution and assume violation of the multivariate normality, she would use the profile-likelihood or MC-ML method for significance testing, but the percentile

bootstrap 95% CI [0.23, 0.90] for the interval estimate. In this case, it appears that the indirect effect ranges from .24 to 0.87. Note that the percentile bootstrap CI is wider than either profile-likelihood or MC-ML method because it is nonparametric and thus does not assume a specific distribution about the data or the residuals.

Discussion

We conducted a large-scale, comprehensive simulation study to evaluate the Type I error rate, statistical power, and coverage of 10 emerging and existing confidence/credible intervals to test an indirect effect in a two-mediator sequential model: (a) Bayesian credible interval with flat prior (Bayes-Flat), (b) Bayesian credible interval with weakly informative prior (Bayes-Weak), (c) Monte Carlo CI with the ML standard errors (MC-ML), (d) Monte Carlo CI with robust Huber-White (Huber, 1967; White, 1980) standard errors (MC-Robust), (e) Bollen and Stine (BS) bootstrap, (f) Yuan, Hayashi, and Yanagihara (YHY) bootstrap, (g) profile likelihood, (h) percentile bootstrap,

(i) bias-corrected (BC) bootstrap, and (j) bias-corrected and accelerated (BCa) bootstrap. A wide range of conditions including sample sizes, size of regression coefficients, and multivariate normal and nonnormal data based on our survey of the published literature were examined in the Monte Carlo simulation study.

For ideal conditions, when the data had a multivariate normal distribution, one key finding was that when $N=50$, the profile likelihood and both Bayesian methods showed the best coverage and accurate Type I error rates while for $N \geq 100$ all methods except for BC and BCa bootstrap showed comparable performance. Popular BC and BCa bootstrap methods frequently showed under-coverage and inflated Type I error rates for tests of indirect effects, especially for smaller sample sizes. The BCa bootstrap reached the maximum Type I error rate of 12.6% when $\beta_1 = \beta_2 = 0.48$, $\beta_3 = 0$, and $N=50$ while the BC bootstrap showed the maximum Type I error rate of 9.8% when $\beta_1 = 0.48$, $\beta_2 = 0.62$, $\beta_3 = 0$, and $N=50$. The lowest coverage for BC and BCa was 87.4%, which occurred when $\beta_1 = 0.36$, $\beta_2 = 0.62$, $\beta_3 = 0.14$, and $N=50$ for BC and $\beta_1 = \beta_2 = \beta_3 = 0.14$, and $N=100$ for BCa. All methods except for BC and BCa showed comparable power across conditions. We also considered two multivariate nonnormality conditions: moderate multivariate nonnormality (skewness = 2, kurtosis = 7) and extreme multivariate nonnormality (skewness = 3, kurtosis = 7). For both multivariate nonnormality conditions, profile-likelihood method, MC-ML, Bayes-Flat, and Bayes-Weak showed the most accurate Type I error rate followed by MC-Robust, percentile bootstrap, YHY, and BC, which showed multiple instances of inflated Type I error rate; power for all methods was comparable across conditions. In terms of coverage, however, the best performing methods were the percentile bootstrap, YHY, and BS method, followed by BC, BCa, and MC-Robust method. The profile-likelihood, MC-ML, and both Bayesian methods showed under-coverage for moderate multivariate nonnormality condition that worsened for extreme multivariate nonnormality condition.

The results of our simulation study complement previous studies. Because of the inflated Type I error rates and under-coverage, we do not recommend BC and BCa. Our recommendation is consistent with the recent studies of BC and BCa for a single-mediator model (Biesanz et al., 2010; Falk, 2018; Falk & Biesanz, 2014; Hayes & Scharkow, 2013). For example, Hayes and Scharkow's (2013) recommended the percentile bootstrap as a compromise between liberal BC bootstrap when mediation occurs and the MC-ML or

the distribution-of-the-product approach (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002) when mediation does not occur. Of course, knowing when mediation occurs or not is not straightforward. However, our recommendation is inconsistent with the recommendation from the previous studies of sequential two-mediator model. Williams and MacKinnon (2008) and Taylor et al. (2008) studied the BC bootstrap CI for normally distributed data, and recommended BC bootstrap; however, these two studies averaged the Type I error rate and coverage across several conditions. The Type I error rates and coverage for specific combination of factors (that were not averaged or aggregated), such as the maximum Type I error rates and the minimum coverage rate, were not reported. Thus, the severity of the inflated Type I error rate and under-coverage rate were not fully explored. For example, for $N=50$, Taylor et al. reported the highest mean (average across conditions) Type I error rate for BC to be .074, which is within Bradley's interval, while in our simulation study the maximum Type I error rate for $N=50$ was .098, which is outside of Bradley's interval. Similarly, when only one of the coefficients was zero, Williams and MacKinnon reported mean Type I error rates, averaged across coefficient values, to be .08093 ($N=50$), .07820 ($N=100$), and .07860 ($N=200$). By comparison, the maximum Type I error rates in our simulation study were .098 ($N=50$), .092 ($N=100$), and .089 ($N=200$). Had we used the average Type I error rate, then BC and BCa would have shown accurate Type I errors across sample sizes. However, the average Type I error rates mask the severity of inflation of the Type I error rates as well as frequency of the inflation of the Type I error rates, as shown in Figure 2.

One important consideration regarding the Bayesian methods considered in the current study is that both likelihood and priors were based on multivariate/univariate normal distributions. Although not in the context of mediation models, according to Zhang (2016), one possible reason that the Bayesian methods did not perform well under multivariate nonnormality conditions is that both likelihood and prior distributions were multivariate/univariate normal distributions. Zhang showed that using a nonnormal distribution to model error terms in latent growth curve models would improve efficiency of standard error estimates for the model parameters. Future studies are needed to further study the effect of using univariate/multivariate nonnormal distributions on the performance of the Bayesian credible intervals for indirect effect.

Unlike previous studies of tests of indirect effect in mediation models, our recommendation is based on the purpose (significance testing, reporting an interval estimate, and reporting a model fit, although model fit was not included in our simulation study) of the mediation analysis and the assumption about the distribution of data. If multivariate normality can be assumed, then one may use (a) profile likelihood, (b) MC-ML, (c) Bayes-Flat, or (d) Bayes-Weak to compute a CI and to conduct significance testing. For multivariate nonnormality conditions, however, we recommend that researchers use different methods to conduct significance testing and to report a CI. For significance testing, we recommend the (a) profile likelihood, (b) MC-ML, (c) Bayes-Flat, or (d) Bayes-Weak method. If the goal is to build a CI without testing for a model fit, then we recommend the percentile bootstrap. Moreover, if one would like to be practical in not choosing the best method in a certain condition but seeking a compromise in choosing only one method for both significance testing and computing a CI, regardless of the distribution of the data, we recommend the percentile bootstrap CI. The percentile bootstrap offers overall accurate (not the most accurate) Type I error, comparable power, and good coverage across conditions. If the researcher's goal is to build a CI for the fit indices that are based on likelihood-ratio chi-squared tests, we recommend using the YHY bootstrap, or with a caveat, the BS bootstrap. Our caveat is that for smaller sample sizes ($N < 200$), BS appeared to produce large standard errors for the coefficients under nonnormality conditions, and many fitted models had convergence problems (Nevitt & Hancock, 2001); note that, however, our simulation study did not find these problems with the two-mediator sequential mediation model. On the other hand, YHY has been recommended for testing a model fit (Zhang & Savalei, 2016). To illustrate the application of the recommended methods we presented an empirical example from a study by Sanchez et al. (2017) whose materials and data are publicly available at (<https://osf.io/g5fvw/>). We provided code for mediation analysis of the example in the supplemental materials.

We made several simplifying assumptions in designing our simulation studies. First, we considered a single sequential mediation chain. However, the sequential mediation chain could be part of a larger model with inclusion of covariates. Inclusion of the additional covariates may improve the estimates of the parameters and their standard errors needed to calculate a CI for an indirect effect. The results of

our simulation study are still applicable to such models, that is, the models with covariates and with non-zero β_4 , β_5 , and β_6 paths in Figure 1. For such models one could, for example, use semi-partial R^2 for endogenous variables and then look up corresponding power in the tabulated power results. We also assumed that the variables, most importantly the mediators, were measured without error. Although this is a common assumption, in practice it will often be violated. In the two-mediator sequential models considered here, measurement errors could attenuate (decrease) the magnitude of indirect effects and inflate (increase) their standard errors (Cohen et al., 2003). We surmise that the results of our simulation studies hold for structural equation models with latent variables used to model measurement errors, assuming the model is correctly specified and appropriately fitting. In addition, we used Vale and Maurelli's (1983) approach to generate multivariate nonnormal data. Vale and Maurelli's approach has been criticized in the literature for underestimating values of skewness and kurtosis in smaller sample sizes (Astivia & Zumbo, 2015), and thus our recommendations regarding the values of skewness and kurtosis should be considered with some degree of caution (particularly in small samples). Nevertheless, the Vale and Maurelli's approach has been widely used in methodological research, and to the extent the method may not produce perfectly defined levels of nonnormality, we are confident that our conclusions hold. Finally, we assumed that the population model in Figure 1 is correctly specified. This assumption implies that the model correctly represents the true causal order of the variables, there are no omitted confounders in the model, and the functional form of the causal relationships is linear. The validity of such assumptions in practice is unknown. Researchers should make attempts to evaluate the effects of violations of the assumptions on their results (Cox et al., 2013; Holland, 1988; Imai et al., 2010; MacKinnon & Pirlott, 2015; Tofighi et al., 2019; Tofighi & Kelley, 2016).

A strength of the simulation studies was that we studied both frequentist and two Bayesian intervals (Bayes-Flat and Bayes-Weak) across a wide range of effect sizes as well as multivariate normality and nonnormality conditions for a more complex mediation model that has two sequential mediators compared to a single mediator. We hope our work helps researchers make more informed decisions regarding how to test for sequential mediation, a method that is

becoming more important in psychology and related disciplines.

Article Information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.


Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was partially supported by NIAAA (R01AA025539, D. Tofighi and K. Witkiewitz, PIs).

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Benjamin B. Dunford, Krannert School of Management, Purdue University, and Katie Witkiewitz, Department of Psychology, University of New Mexico, for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

ORCID

Davood Tofighi  <http://orcid.org/0000-0001-8523-7776>
Ken Kelley  <http://orcid.org/0000-0002-4756-8360>

References

- Adamczyk, K. (2018). Direct and indirect effects of relationship status through unmet need to belong and fear of being single on young adults' romantic loneliness. *Personality and Individual Differences, 124*, 124–129. doi:10.1016/j.paid.2017.12.011
- Andreassen, T. W., Lorentzen, B. G., & Olsson, U. H. (2006). The impact of non-normality and estimation methods in SEM on satisfaction research in marketing. *Quality & Quantity, 40*, 39–58. doi:10.1007/s11135-005-4510-y
- Astivia, O. L. O., & Zumbo, B. D. (2015). A cautionary note on the use of the Vale and Maurelli method to generate multivariate, nonnormal data for simulation purposes. *Educational and Psychological Measurement, 75*(4), 541–567. doi:10.1177/0013164414548894
- Ato García, M., Vallejo Seco, G., & Ato Lozano, E. (2014). Classical and causal inference approaches to statistical mediation analysis. *Psicothema, 26*(2), 252–259. <https://doi.org/10.7334/psicothema2013.314>
- Bernier, A., McMahon, C. A., & Perrier, R. (2017). Maternal mind-mindedness and children's school readiness: A longitudinal study of developmental processes. *Developmental Psychology, 53*(2), 210–221. doi:10.1037/dev0000225
- Biesanz, J. C., Falk, C. F., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research, 45*(4), 661–701. doi:10.1080/00273171.2010.498292
- Bollen, K. A., & Stine, R. A. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology, 20*, 115–140. doi:10.2307/271084
- Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21*, 205–229. doi:10.1177/0049124192021002004
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x
- Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods, 49*(5), 1716–1735. doi:10.3758/s13428-016-0814-1
- Canty, A., & Ripley, B. D. (2017). *Boot: Bootstrap R (S-Plus) Functions (Version 1.3-20)*. [Computer software]. Available from <https://cran.r-project.org/web/packages/boot/index.html>
- Chen, J., Choi, J., Weiss, B. A., & Stapleton, L. (2014). An empirical evaluation of mediation effect analysis with manifest and latent variables using Markov Chain Monte Carlo and alternative estimation methods. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(2), 253–262. doi:10.1080/10705511.2014.882688
- Chernick, M. R., & LaBudde, R. A. (2011). *An introduction to bootstrap methods with applications to R*. Hoboken, NJ: Wiley.
- Cheung, M. W. L. (2007). Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(2), 227–246. doi:10.1080/10705510709336745
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cox, D. R., & Hinkley, D. V. (2000). *Theoretical statistics*. Boca Raton, FL: Chapman & Hall/CRC.
- Cox, M. G., Kisbu-Sakarya, Y., Miočević, M., & MacKinnon, D. P. (2013). Sensitivity plots for confounder bias in the single mediator model. *Evaluation Review, 37*(5), 405–431. doi:10.1177/0193841X14524576
- Craig, C. C. (1936). On the frequency function of xy . *The Annals of Mathematical Statistics, 7*(1), 1–15. doi:10.1214/aoms/1177732541
- 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

- Davison, A. C. (2003). *Statistical models*. Cambridge, UK: Cambridge University Press.
- Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their application*. New York, NY: Cambridge University Press.
- Deković, M., Asscher, J. J., Manders, W. A., Prins, P. J. M., & van der Laan, P. (2012). Within-intervention change: Mediators of intervention effects during multisystemic therapy. *Journal of Consulting and Clinical Psychology*, 80(4), 574–587. doi:10.1037/a0028482
- Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2), 216–222. doi:10.1016/0370-2693(87)91197-X
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185. doi:10.2307/2289144
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.
- Falk, C. F. (2018). Are robust standard errors the best approach for interval estimation with nonnormal data in structural equation modeling? *Structural Equation Modeling: A Multidisciplinary Journal*, 25(2), 244–266. doi:10.1080/10705511.2017.1367254
- Falk, C. F., & Biesanz, J. C. (2015). Inference and interval estimation methods for indirect effects with latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, 22, 24–38. doi: 10.1080/10705511.2014.935266
- Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(2), 87–107. doi:10.1080/10705519709540063
- Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Los Angeles, CA: SAGE.
- Freedman, D. A. (2006). On the so-called “Huber Sandwich Estimator” and “Robust Standard Errors.”. *The American Statistician*, 60(4), 299–302. doi:10.1198/000313006X152207
- Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research*, 47(1), 61–87. doi:10.1080/00273171.2012.640596
- Gabry, J., & Goodrich, B. (2018, April 13). Prior distributions for rstanarm models. Retrieved September 26, 2018, from Prior distributions for rstanarm models website: <http://mc-stan.org/rstanarm/articles/priors.html#how-to-specify-flat-priors-and-why-you-typically-shouldnt>
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721–741. doi:10.1109/TPAMI.1984.4767596
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1998). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall.
- Graham, J. D., Martin Ginis, K. A., & Bray, S. R. (2017). Exertion of self-control increases fatigue, reduces task self-efficacy, and impairs performance of resistance exercise. *Sport, Exercise, and Performance Psychology*, 6(1), 70–88. doi:10.1037/spy0000074
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109. doi:10.1093/biomet/57.1.97
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: The Guilford Press.
- Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis does method really matter? *Psychological Science*, 24(10), 1918–1927. doi:10.1177/0956797613480187
- Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 19(10), 3595–3617. doi: 10.1080/03610929008830400
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18, 449–484. doi:10.2307/271055
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 221–233. Berkeley, CA: University of California Press.
- Huertas-Valdivia, I., Llorens-Montes, F. J., & Ruiz-Moreno, A. (2018). Achieving engagement among hospitality employees: A serial mediation model. *International Journal of Contemporary Hospitality Management*, 30(1), 217–241. doi:10.1108/IJCHM-09-2016-0538
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309–334. doi:10.1037/a0020761
- Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, 117(2), 348–357. doi:10.1037/0033-2909.117.2.348
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.
- Koning, I. M., Maric, M., MacKinnon, D., & Vollebergh, W. A. M. (2015). Effects of a combined parent–student alcohol prevention program on intermediate factors and adolescents’ drinking behavior: A sequential mediation model. *Journal of Consulting and Clinical Psychology*, 83(4), 719–727. doi:10.1037/a0039197
- Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York, NY: Springer.
- MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 51(1), 201–226. doi: 10.1146/annurev.psych.51.1.201
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Erlbaum.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1), 83–104. doi:10.1037//1082-989X.7.1.83
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39(1), 99–128. doi:10.1207/s15327906mbr3901_4

- MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, 19(1), 30–43. doi:10.1177/1088868314542878
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530. doi:10.1093/biomet/57.3.519
- McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–126). Orlando, FL: Academic Press.
- McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.
- Mecklin, C. J., & Mundfrom, D. J. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, 75(2), 93–107. doi:10.1080/0094965042000193233
- Meeker, W. Q., & Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, 49, 48. doi:10.2307/2684811
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247), 335–341. doi:10.2307/2280232
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156–166. doi:10.1037/0033-2909.105.1.156
- Muth, C., Oravecz, Z., & Gabry, J. (2018). User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. *The Quantitative Methods for Psychology*, 14(2), 99–119. doi:10.20982/tqmp.14.2.p099
- Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335. doi:10.1037/a0026802
- Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 113–162). Boca Raton, FL: Chapman & Hall/CRC.
- Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., ... Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, 81(2), 535–549. doi:10.1007/s11336-014-9435-8
- Neale, M. C., & Miller, M. B. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics*, 27(2), 113–120.
- Nevitt, J., & Hancock, G. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(3), 353–377. doi:10.1207/S15328007SEM0803_2
- Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(4), 557–595. doi:10.1207/S15328007SEM0704_3
- Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. New York, NY: Oxford University Press.
- Pek, J., & Wu, H. (2015). Profile likelihood-based confidence intervals and regions for structural equation models. *Psychometrika*, 80(4), 1123–1145. doi:10.1007/s11336-015-9461-1
- Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, 67(4), 741–763. doi:10.1037//0022-3514.67.4.741
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3), 879–891. doi:10.3758/BRM.40.3.879
- Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, 6(2), 77–98. doi:10.1080/19312458.2012.679848
- Reh, S., Tröster, C., & Van Quaquebeke, N. (2018). Keeping (future) rivals down: Temporal social comparison predicts coworker undermining via future status threat and envy. *Journal of Applied Psychology*, 103(4), 399–415. doi:10.1037/apl0000281
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.
- Sanchez, D. T., Chaney, K. E., Manuel, S. K., Wilton, L. S., & Remedios, J. D. (2017). Stigma by prejudice transfer: Racism threatens White women and sexism threatens men of color. *Psychological Science*, 28(4), 445–461. doi:10.1177/0956797616686218
- Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, 75(2), 243–248. doi:10.1007/s11336-009-9135-y
- Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), 149–160. doi:10.1080/10705511.2013.824793
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7(4), 422–445. doi:10.1037//1082-989X.7.4.422
- Springer, M. D., & Thompson, W. E. (1966). The distribution of products of independent random variables. *SIAM Journal on Applied Mathematics*, 14(3), 511–526. doi:10.1137/0114046
- Stan Development Team. (2018). rstanarm: Bayesian applied regression modeling via Stan (Version 2.17.4). Retrieved from <http://mc-stan.org/>
- Taylor, A. B., MacKinnon, D. P., & Tein, J.-Y. (2008). Tests of the three-path mediated effect. *Organizational Research Methods*, 11(2), 241–269. doi:10.1177/1094428107300344
- Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(3), 510–534. doi:10.1080/10705511.2010.489379

- Tofighi, D., Hsiao, Y.-Y., Kruger, E. S., MacKinnon, D. P., Van Horn, M. L., & Witkiewitz, K. (2019). Sensitivity analysis of the no-omitted confounder assumption in latent growth curve mediation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(1), 94–109. <https://doi.org/10.1080/10705511.2018.1506925>
- Tofighi, D., & Kelley, K. (2016). Assessing omitted confounder bias in multilevel mediation models. *Multivariate Behavioral Research*, 51(1), 86–105. <https://doi.org/10.1080/00273171.2015.1105736>
- Tofighi, D., & MacKinnon, D. P. (2016). Monte Carlo confidence intervals for complex functions of indirect effects. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(2), 194–205. <https://doi.org/10.1080/10705511.2015.1057284>
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48(3), 465–471. doi:10.1007/BF02293687
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York, NY: Oxford University Press.
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: SAGE.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838. doi:10.2307/1912934
- Williams, J., & MacKinnon, D. P. (2008). Resampling and distribution of the product methods for testing indirect effects in complex models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1), 23–51. doi:10.1080/10705510701758166
- Yuan, K.-H., Hayashi, K., & Yanagihara, H. (2007). A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research*, 42(2), 261–281. doi:10.1080/00273170701360662
- Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, 14(4), 301–322. doi:10.1037/a0016972
- Zhang, X., & Savalei, V. (2016). Bootstrapping confidence intervals for fit indexes in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3), 392–408. doi:10.1080/10705511.2015.1118692
- Zhang, Z. (2016). Modeling error distributions of growth curve models through Bayesian methods. *Behavior Research Methods*, 48(2), 427–444. doi:10.3758/s13428-015-0589-9