7-2018

A Comparison of Alternative Bias-Corrections in the Bias-Corrected Bootstrap Test of Mediation

Donna Chen
University of Nebraska-Lincoln, donna.chen59@gmail.com

A COMPARISON OF ALTERNATIVE BIAS-CORRECTIONS IN THE BIAS-CORRECTED BOOTSTRAP TEST OF MEDIATION

by

Donna Chen

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor Matthew S. Fritz

Lincoln, Nebraska

August, 2018

A COMPARISON OF ALTERNATIVE BIAS-CORRECTIONS IN THE BIAS-CORRECTED BOOTSTRAP TEST OF MEDIATION

Donna Chen, M.A.

University of Nebraska, 2018

Advisor: Matthew S. Fritz

Although the bias-corrected (BC) bootstrap is an oft recommended method for obtaining more powerful confidence intervals in mediation analysis, it has also been found to have elevated Type I error rates in conditions with small sample sizes. Given that the BC bootstrap is used most often in studies with low power due to small sample size, the focus of this study is to consider alternative measures of bias that will reduce the elevated Type I error rate without reducing power. The alternatives examined fall under two categories: bias correction and transformation. Although the bias correction methods did not significantly decrease Type I error rate, the associated confidence intervals were similar to the original BC bootstrap. The transformations, however, did not produce confidence intervals with more accurate Type I error rate.

For my parents, my sisters, and my friends. Your support and encouragement mean the world to me. Thank you.

Acknowledgments

A huge thank you to Matthew Fritz for his guidance, mentorship, teaching, and support. Thank you also to my committee members, Lorey Wheeler and Jim Bovaird, for their invaluable insights and feedback. Special shout out to Jayden Nord for acting as a sounding board and laying the foundations for efficient programming.

Table of Contents

List of Figures

List of Tables

A Comparison of Alternative Bias-Corrections in the Bias-Corrected Bootstrap Test of Mediation

Experimental research studies are conducted in an attempt to understand the causal relationships between variables. In the social sciences, however, causal relationships are not always direct. Instead, causal relationships may involve one or more intermediate variables, known as mediators. The sampling distribution for these indirect causal effects are often not normally distributed, meaning the use of traditional normal-theory statistical tests for mediators, known as statistical mediation analysis, may result in incorrect inferences (Lomnicki, 1967; Springer & Thompson, 1966). One solution is to use resampling methods that do not make a normality assumption, such as the bootstrap (Efron & Tibshirani, 1993).

Many researchers have recommended using the bootstrap for mediation analysis (e.g., Shrout & Bolger, 2002), including MacKinnon, Lockwood, and Williams (2004) who compared the performance of multiple variations of the bootstrap for testing mediated effects, such as the percentile, bias-corrected, and accelerated bias-corrected bootstrap tests. They found that the bias-corrected bootstrap had the highest statistical power but also elevated Type I error rates in certain conditions. In a follow-up study, Fritz, Taylor, and MacKinnon (2012) found the Type I error rates for the bias-corrected bootstrap were most inflated when the sample size was small. Given that the bias-corrected bootstrap is most likely to be used in studies with low statistical power due to small sample sizes, the exact situation where the Type I error rates for the bias-corrected bootstrap are the worst, there is a need to determine if the bias-corrected bootstrap can be modified in order to reduce the Type I error while maintaining the increased statistical power. Therefore, the purpose of this study is to identify and compare alternative corrections for bias in bootstrap tests of mediation.

Literature Review

Differentiating Mediation and Moderation

The terms *mediation* and *moderation* are often incorrectly used synonymously. Baron and Kenny's (1986) widely cited distinction between moderation and mediation is that a *moderator* "affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable" (p. 1174) and a *mediator* "accounts for the relation between the predictor and the criterion" (p. 1176). Moderation examines the *when*, mediation defines the *how* or *why*. Another way to conceptualize the difference is that moderation refers to the *interaction effect* between two variables (e.g., "At which level does the effect occur?" or "For which group of people is the treatment effective?"), while mediation is an *indirect effect* on the dependent variable through an intervening variable. This study focuses on issues of statistical mediation analysis, and we begin by discussing the single-mediator model.

The Single-Mediator Model

A total effect model involves the independent variable (X) causing the dependent variable (Y), as illustrated in Figure 1.

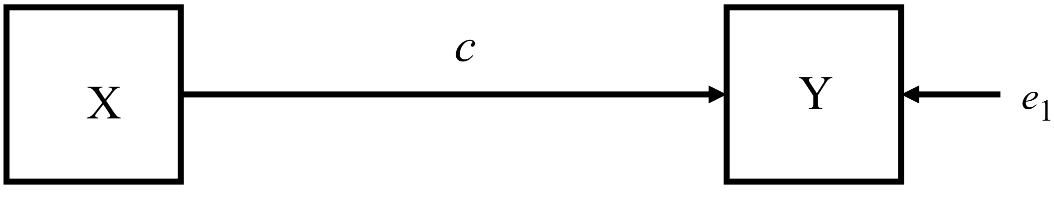

Figure 1. Path diagram for the total effect model.

The total effect model can be represented by the regression equation

$$\mathrm{Y_i} = i_1 + c\mathrm{X_i} + e_1 \tag{1}$$

where i_1 is the intercept, c is the effect of X on Y, and e_1 is the part of Y that is not explained by the relation between X and Y.

Mediation analysis examines the effect of an intermediate variable within a causal sequence. Mediation occurs when the mediator (M) comes between X and Y so that X is the effect on M, and M is the effect on Y, as illustrated in Figure 2. Other common terminology used to describe X and Y are antecedent variables (X) and consequent variables (Y).

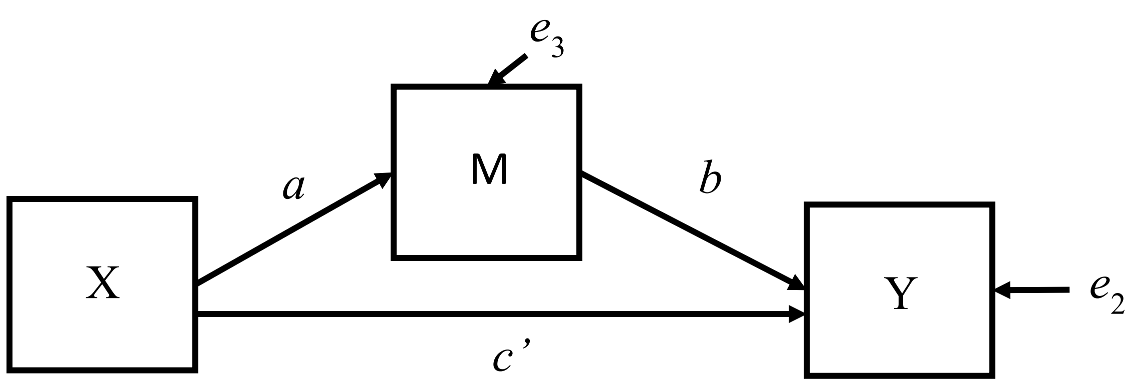

Figure 2. Path diagram for a mediation model.

The mediation model can be represented by the following regression equations:

$$\mathrm{Y_i} = i_2 + c'\mathrm{X_i} + bM_i + e_2 \tag{2}$$

$$\mathrm{M_i} = i_3 + a\mathrm{X_i} + \ e_3 \tag{3}$$

In these two equations, i_2 and i_3 are the intercepts, a is the relation between X and M, b is the relation between M and Y adjusted for the effects of X, and e_2 and e_3 are the unexplained or error variability.

In the presence of a mediator, the X to M to Y relationship can be referred to in three different ways: the indirect effect, the mediated effect, or the quantity ab. The quantity ab describes the amount of change between X and Y indirectly through M, which is equivalent to saying it is the effect of X on M (parameter a) multiplied by the effect of M on Y (parameter b). On the other hand, c' is the direct effect of X on Y, accounting for the mediator effect. Another way to calculate the mediated effect is by using $ab = c - c'$ because the difference between the total effect and the direct effect is the indirect effect (Judd & Kenny, 1981). In a sample estimate, however, the notation for a, b, c, and c' would be $\hat{a},\ \hat{b},\ \hat{c},\ \text{and}\ \widehat{c'}$, respectively, where ˆ denotes estimates of the parameters.

Assumptions of Mediation Analysis

Mediation analysis is based on a number of assumptions, where violating the assumptions may result in incorrect interpretation of the data. MacKinnon (2008) lists the assumptions for the mediation regression equations and the single-mediator model.

1. **Correct functional form.** Equations 1 thru 3 assume that the variables have a linear relationship; when the independent variable changes by 1 unit, the dependent variable also changes in a specified amount. For example, when X changes by 1 unit, there are $\hat{a}$ units change in M. Equations 1 thru 3, however, do not always have to be used, in which case this assumption is not made. For example, if the relations among variables are nonlinear, the correct transformations are made to reflect a nonlinear relation. Additionally, the variables are additive; there is no interaction between variables.

2. **No omitted influences.** The three variables in the single-mediator model represent the true underlying model, and there are no other variables that affect or are related to these three variables.
3. **Accurate measurement.** The measures used to examine variables X, M, and Y are reliable and valid.
4. **Well-behaved residuals.** The errors, or residuals, for one observation are not correlated with errors from other observations. Error variances are also constant at each predictor value.
5. **Normally distributed X, M, and Y.** Each of these variables has a normal distribution.
6. **Temporal precedence.** There is an ordering of the variables over time, such that X comes before M, and M comes before Y.
7. **Measurement timing.** There is a true timing of change between the independent variable, mediator, and dependent variable, and the mediational chain being measured is representative of this true timing.
8. **Micro versus macro mediational chain.** In relation to temporal precedence, a single-mediator model may be derived from a mediational chain which consists of many links or steps. The researcher's job is to distinguish the macromediational from the micromediational chain and determine which steps to measure accordingly. Given restrictions on resources and measurement of only a single-mediator, however, it may be that only a small section of a long mediational chain is measured. Thus, the results may be that either the correct

steps in the micromediational chain are measured or a real mediation effect is missed.

9. **Theoretical versus empirical mediator.** The possibility of a statistically significant mediated effect may not be the true effect and may instead be a proxy for another mediator.

Tests of the Mediated Effect

Judd and Kenny (1981) originally proposed a series of regression tests in order to detect mediation when three conditions are met. The first condition states that there must be an effect of treatment on the outcome variable. Second, in a causal chain, all variables affect the variable that follows when all previous variables are controlled for. Third, if the mediator is controlled, the treatment has no effect on the outcome variable. Baron and Kenny (1986) expanded Judd and Kenny's (1981) method by introducing more leniency towards partial mediation (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). Baron and Kenny illustrate how each of the causal effects are individually tested for significance, rendering hierarchical and stepwise regression unnecessary. In order to establish a mediation effect, the three regression models (represented by Equations 1 thru 3) should be estimated.

MacKinnon (2008) frames the causal steps method in terms of the estimated regression coefficients $\hat{a}$, $\hat{b}$, $\hat{c}$, and $\widehat{c'}$:

1. The coefficient $\hat{c}$, which tests the relation between X and Y, must be statistically significant. If it is not, we assume there is no mediated effect and the tests end here.

2. The coefficient $\hat{a}$, which tests the relation between X and M, must be statistically significant.
3. The coefficient $\hat{b}$, which tests the relation between M and Y when X is controlled, must be statistically significant.
4. The coefficient $\widehat{c'}$, the direct effect, must be nonsignificant. This step does not take into account partial mediation, so recent changes have been made to require that $\widehat{c'} < \hat{c}$.

Since the development of the causal steps approach, alternative methods for detecting mediated effects have been proposed. MacKinnon et al. (2002) compared fourteen methods used to test for the statistical significance of mediation effects. The authors separated these approaches into three categories: causal steps, difference in coefficients, and the product of coefficients. The causal steps approach is consistent with Judd and Kenny (1981) and Baron and Kenny's (1986) proposed methods. In the second category, the difference in coefficients tests cover a wider set of relations between the independent and dependent variables to assess the effect of the intervening variable. Different pairs of coefficients can be compared, such as regression coefficients ($\hat{c} - \widehat{c'}$) or correlation coefficients. The coefficients tests give an estimate of the intervening variable effect along with its standard error. The third category is the product of coefficients method which takes the indirect effect, $\hat{a}\hat{b}$, and divides it by its standard error. Both the difference in coefficients tests and the product of coefficients tests compare the estimate of the intervening variable effect to a known sampling distribution (Fritz & MacKinnon, 2007; MacKinnon et al., 2002).

Fritz and MacKinnon (2007) and MacKinnon et al. (2002) have demonstrated that the causal steps approach, while commonly used, is often underpowered. MacKinnon (2008) further notes that in order to establish mediation, only steps 2 and 3 need to be statistically significant. If $\hat{a}$ and $\hat{b}$ are both found to be significant, mediation is present. This is called the joint significance test. An issue with both the difference in coefficients tests and the product of coefficients tests is their basis on the assumption of normally distributed random variables, which is not the case often observed in mediation analysis. Additionally, these tests do not actually test for *ab*, which becomes problematic when testing more complex models with multiple mediators.

Calculating Confidence Intervals for a Mediated Effect

Alternatively, MacKinnon (2008) suggests confidence intervals be constructed to statistically test for mediated effects. Confidence intervals allow researchers to provide a range of possible values for the effect so the result is neither a single value nor a push for researchers to make a definite choice between 'reject' or 'fail to reject'. There are two ways to categorize confidence intervals: parametric versus nonparametric.

Parametric Methods

One way to calculate confidence intervals is by assuming the data are distributed normally and using the z-distribution such that the confidence intervals are calculated using

$$CI: \hat{a}\hat{b} \pm z_{Type\ I\ error} * \sigma_{\hat{a}\hat{b}} \tag{4}$$

where $\hat{a}\hat{b}$ is the mediated effect, $z_{Type\ I\ error}$ is the z critical value on a standard normal distribution with the specified α Type I error rate, and $\sigma_{\hat{a}\hat{b}}$ is the standard error of the mediated effect.

Indirect effects, however, have been shown to be skewed, so asymmetric confidence limits are calculated instead (MacKinnon et al., 2004). MacKinnon, Fritz, Williams, and Lockwood (2007) developed a test involving the calculation of confidence intervals with a Fortran program known as PRODCLIN. The program uses standardized values of $\hat{\alpha}$, $\widehat{\beta}$, and Type I error rate to automatically calculate asymmetric confidence intervals. If zero falls within a confidence interval, the mediated effect is not significantly different from zero and is therefore not statistically significant. Another way to create asymmetric confidence limits is through nonparametric resampling methods.

Nonparametric Methods

Parametric methods (e.g., normal theory and PRODCLIN) are based on the assumption that the data were collected from a known distribution where the form of the population distribution is “completely specified except for a finite number of parameters” (Higgins, 2004, p. 7). Higgins explains that data analysis often begins by ascertaining the fit of a normal distribution to the data. The data, however, do not always come from a normal distribution, nor do they always meet parametric assumptions. This is the issue with estimating indirect effects in mediation analysis; the distribution is observed to be skewed. Contrary to parametric methods, nonparametric methods (e.g., permutations and resampling methods) are considered for analyzing non-normal data because they operate under minimal assumptions about distribution form.

Permutations. A large class of nonparametric methods is categorized as permutations. Permutation tests are based on calculating sample statistics from all possible combinations of data randomly sorted to each of the treatment groups. Given *m* units in treatment 1 and *n* units in treatment 2, the number of observed data units would be *m* + *n*. The number of possible permutations equals

$$\binom{m+n}{m} = \frac{(m+n)!}{m!n!} \tag{5}$$

The problem with the permutation test is that the number of possible replications increases rapidly. For example, when *m* = 3 and *n* = 2, the number of possible permutations equals 10. However, when *m* = 10 and *n* = 15, there would be 3,268,760 possible permutations. The number of possible permutations quickly escalates, and the possibility of obtaining all possible replications for a study will eventually become tedious or impossible. The bootstrap method reduces the number of samples and is therefore an alternative to permutation tests.

Bootstrapping. In 1979, Bradley Efron proposed the bootstrap as a nonparametric method that samples with replacement. Its name is derived from the idea of "pulling oneself up by one's bootstrap" (Efron & Tibshirani, 1993, p. 5). The power of the modern computer allows statisticians to use bootstrapping as a means for making certain statistical inferences and to estimate the accuracy of the sample statistics in relation to the population parameter (Higgins, 2004). Bootstrap sampling is an application of the plug-in principle, which is a method of using the sample to estimate parameters for the population (Efron & Tibshirani, 1993).

Bootstrapping treats the existing sample data as the population, wherein samples are taken out of the existing dataset as if to sample from the population (Higgins, 2004). Given a sample size *n*, a new sample size *n* will be drawn from the existing sample, and statistics will be calculated. This process is repeated a large amount of times (usually upwards of 1000 replications) and the distribution for the bootstrap samples is formed.

Each time a number is drawn, it is returned to the sample so that each number has an equal chance of being selected again (replacement). Therefore, there is a possibility of drawing any one specific number 0, 1, 2, 3,…*n* times in the bootstrap sample.

Mediation and Bootstrap

In considering statistical mediation analysis, a bootstrap sample is drawn from the original data and the mediated effect is calculated. Calculating the mediated effect is repeated for each bootstrap sample (e.g., 1,000) and the estimates are used to form a distribution. There are many different flavors of bootstrapping, but we discuss three in particular that are used to calculate confidence intervals: the percentile bootstrap, the bias-corrected bootstrap, and the accelerated bias-corrected bootstrap.

Percentile bootstrap. The simplest form of bootstrapping is the percentile bootstrap where the bootstrap samples are ordered from smallest to largest, and the exact percentiles corresponding to the set alpha level are used as the upper and lower bounds. Instead of using a *z*-score from a standard normal table, bootstrapping constructs a new distribution around the mediated effect and takes the percentiles from the new distribution. For example, given 1000 bootstrap samples, a 95% confidence interval is calculated by finding the exact values at the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles such that the new confidence interval is

$$\hat{a}\hat{b}^{*(.025)} \leq \hat{a}\hat{b} \leq \hat{a}\hat{b}^{*(.975)} \tag{6}$$

where $\hat{a}\hat{b}^{*(.025)}$ and $\hat{a}\hat{b}^{*(.975)}$ are the 25$^{\text{th}}$ and 975$^{\text{th}}$ values on the bootstrapped distribution.

Bias-corrected bootstrap. The bias-corrected (BC) bootstrap also uses percentiles as the lower and upper confidence limits, but these limits may not be the same as the percentile method. Instead, the endpoints of the confidence interval are recalculated using the bias-correction, $\hat{z}_0$. Specific to the mediated effect, $\hat{z}_{\hat{a}\hat{b}}$ will be used to denote the bias-correction. The bias-correction is obtained by calculating the proportion of bootstrap replications of $\hat{a}\hat{b}$, denoted by $\hat{a}\hat{b}^*(b)$, that are less than the original estimate of the mediated effect $\hat{a}\hat{b}$,

$$\hat{z}_{\hat{a}\hat{b}} = \Phi^{-1}\left(\frac{\#\{\hat{a}\hat{b}^*(b) < \hat{a}\hat{b}\}}{B}\right) \tag{7}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of a standard normal cumulative distribution function, B is the number of bootstrap samples, and $\#\{\hat{a}\hat{b}^*(b) < \hat{a}\hat{b}\}$ means the number of bootstrap replications that are less than $\hat{a}\hat{b}$. The proportion is used on the standard normal distribution to find the corresponding z-score. The following equations from Efron and Tibshirani (1993) are used to define the BC confidence limits:

$$\text{BC} : \left(\hat{a}\hat{b}_{lo}, \hat{a}\hat{b}_{up}\right) = (\hat{a}\hat{b}^{*(\alpha_1)}, \hat{a}\hat{b}^{*(\alpha_2)}) \tag{8}$$

where

$$\alpha_1 = \Phi\left(2\hat{z}_{\hat{a}\hat{b}} + z^{(\alpha)}\right) \tag{9}$$

$$\alpha_2 = \Phi\left(2\hat{z}_{\hat{a}\hat{b}} + z^{(1-\alpha)}\right) \tag{10}$$

given that $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\hat{z}_{\hat{a}\hat{b}}$ is the measure of bias for the indirect effect, and $z^{(\alpha)}$ is the 100αth percentile point of a standard normal distribution. Considering $\alpha = .05$, the the lower confidence bound is calculated using $2\hat{z}_{\hat{a}\hat{b}} - 1.96$, and the upper confidence bound is calculated using $2\hat{z}_{\hat{a}\hat{b}} + 1.96$. If the proportion below $\hat{a}\hat{b}$ is .50, the bias-corrected bootstrap is equivalent to the percentile bootstrap.

Accelerated bias-corrected bootstrap. The accelerated bias-corrected (BC_a) bootstrap is an improvement over the percentile bootstrap due to its higher accuracy (Efron & Tibshirani, 1993). The BC_a accounts for two components, the bias-correction ($\hat{z}_0$) and the acceleration ($\hat{k}$). The bias-correction is calculated using equation 7. Acceleration refers to the rate of change in the standard deviation of $\hat{a}\hat{b}$ as ab varies and is calculated by

$$\hat{k} = \frac{\sum_{i=1}^{n}\left(\hat{a}\hat{b}_{(\cdot)} - \hat{a}\hat{b}_{(-i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\hat{a}\hat{b}_{(\cdot)} - \hat{a}\hat{b}_{(-i)}\right)^2\right]^{3/2}} \tag{11}$$

where $\hat{a}\hat{b}_{(-i)}$ is the estimate of ab with case *i* deleted from the original data set, also called the *i*th jackknife estimate of ab, and $\hat{a}\hat{b}_{(\cdot)}$ is the mean of all the jackknife estimates of ab. Equation 8 is also applicable to the BC_a with the difference being in the values of α_1 and α_2 where $\hat{k} \neq 0$:

$$\alpha_1 = \Phi\left(\hat{z}_{\hat{a}\hat{b}} + \frac{\hat{z}_{\hat{a}\hat{b}} + z^{(\alpha)}}{1 - \hat{k}\left(\hat{z}_{\hat{a}\hat{b}} + z^{(\alpha)}\right)}\right) \tag{12}$$

$$\alpha_2 = \Phi\left(\hat{z}_{\hat{a}\hat{b}} + \frac{\hat{z}_{\hat{a}\hat{b}} + z^{(1-\alpha)}}{1 - \hat{k}\left(\hat{z}_{\hat{a}\hat{b}} + z^{(1-\alpha)}\right)}\right). \tag{13}$$

When $\hat{k} = 0$, the BC_a reduces to the BC.

Issues surrounding current bias-corrected bootstrap. Although the bias-corrected bootstrap has been found to have relatively higher statistical power, Type I error rate has also been found to be elevated in certain conditions. Fritz et al. (2012) found that Type I error rates occur as an interaction effect between path size and sample size, “such that elevated Type I error rates occur when the sample size is small and the effect size of the nonzero path is medium or larger” (p. 61). The estimate $\hat{z}_0$ measures the median bias of the bootstrapped sample instead of the mean bias. In referencing the accelerated bias-corrected bootstrap, Efron and Tibshirani note that “it is in fact easier to get a good estimate for ‘a’ than for z_0” (1993, p. 327). Based on this, Fritz et al. (2012) suggest the need to find a better estimate of bias.

Current Study

The following eight bias-correction methods are proposed as alternatives to Efron and Tibsharani’s z_0 that will not have elevated Type I error rates when calculating confidence levels for mediated effects with small samples. The alternatives can be categorized into two groups: measures of bias and transformations.

Alternative Measures of Bias

Median (z_{medi}). Instead of using the estimate of $\hat{a}\hat{b}$, perhaps a measure of central tendency would be a better bias-correction. One alternative is to find the proportion of bootstrap replications that fall below the median of the bootstrap sampling distribution. Considering the proportion that should fall below the median (50%), however, the bias-correction should equal zero and the values for the median confidence interval are expected to equal the percentile bootstrap interval. The confidence interval is equal to

$$\mathrm{BC_{medi}} : (\hat{a}\hat{b}^{*(\alpha_{medi1})}, \hat{a}\hat{b}^{*(\alpha_{medi2})}) \tag{14}$$

where

$$\alpha_{medi1} = \Phi\left(2\hat{z}_{medi} + z^{(\alpha/2)}\right) \tag{15}$$

$$\alpha_{medi2} = \Phi\left(2\hat{z}_{medi} + z^{(1-\alpha/2)}\right) \tag{16}$$

The bias-correction can be calculated by

$$\hat{z}_{medi} = \Phi^{-1}\left(\frac{\#\{\hat{a}\hat{b}^{*}(b) < \hat{a}\hat{b}^{*}_{median}\}}{B}\right). \tag{17}$$

The bias-correction is multiplied by two and added to the 100αth percentile point of a standard normal distribution percentile values at the $\alpha/2$ and the $1-\alpha/2$ level on a standard normal cumulative distribution.

Mean (z_{mean}). The second alternative is to use another measure of central tendency. As an alternative to calculating the proportion of bootstrap replications that fall below the value $\hat{a}\hat{b}$, the proportion of bootstrap replications that fall below the mean of the bootstrap sampling distribution will be calculated. The confidence interval is equal to

$$\mathrm{BC}_{\mathrm{mean}} : (\hat{a}\hat{b}^{*(\alpha_{mean1})}, \hat{a}\hat{b}^{*(\alpha_{mean2})}) \quad (18)$$

where

$$\alpha_{mean1} = \Phi\left(2\hat{z}_{mean} + z^{(\alpha/2)}\right) \quad (19)$$

$$\alpha_{mean2} = \Phi\left(2\hat{z}_{mean} + z^{(1-\alpha/2)}\right) \quad (20)$$

The bias-correction is calculated by

$$\hat{z}_{mean} = \Phi^{-1}\left(\frac{\#\{\hat{a}\hat{b}^{*}(b) < \hat{a}\hat{b}^{*}_{mean}\}}{B}\right) \quad (21)$$

Calculating the confidence interval uses the same method as described above for the median. The only value that changes is the bias-correction; $\hat{z}_{mean}$ is used in place of $\hat{z}_{medi}$.

Traditional measure of sample skewness—$g1$ (z_{g1}). Another way to take bias into account is to consider the skewness of the product distribution. The traditional measure of skewness—g_1 is defined by

$$g_1 = \frac{m_3}{m_2^{3/2}} \quad (22)$$

$$m_r = \frac{1}{n}\Sigma(x_i - \bar{x})^r \quad (23)$$

where *n* denotes sample size (Joanes & Gill, 1998). For the numerator of g_1, the sum of the differences between each score and the mean are raised to the third power and divided by the sample size. The denominator is the sum of the differences between each score and the mean squared, divided by the sample size.

Although the measure of skewness is unbounded, a simulation using the parameters of this study was run to examine the range of possible skewness coefficients, which turned out to be [-3.48, 2.97]. The range of the skewness coefficients was deemed reasonable (e.g., between z-scores of -4 and 4 as opposed to ranging to infinity) so the unbounded nature of g_1 was not of major concern. The measure of skewness becomes the bias-correction so that the confidence interval is equal to

$$\mathrm{BC}_{g1} : (\hat{a}\hat{b}^{*(\alpha_{g11})}, \hat{a}\hat{b}^{*(\alpha_{g12})}) \tag{24}$$

where

$$\alpha_{g11} = \Phi\left(2\hat{z}_{g1} + z^{(\alpha/2)}\right) \tag{25}$$

$$\alpha_{g12} = \Phi\left(2\hat{z}_{g1} + z^{(1-\alpha/2)}\right) \tag{26}$$

The bias-correction is

$$\hat{z}_{g1} = g_1 \tag{27}$$

Given the skewness of the bootstrap sample, the corresponding z-score is used as the bias-correction.

Medcouple (z_{mc}). The medcouple is a robust measure of skewness that “measures the (standardized) difference between the distances of x_j and x_i to the median” (Brys, Hubert, & Struyf, 2004, p. 998). It is more robust towards outliers than the classic measure of skewness. The median, m_n, is usually defined as

$$m_n = \begin{cases} \frac{x_{n/2}+x_{n/2+1}}{2} & \text{if } n \text{ is even} \\ x_{(n+1)/2} & \text{if } n \text{ is odd.} \end{cases} \tag{28}$$

The medcouple, introduced by Brys, Huber, and Struyf (2003) is defined as

$$\mathrm{MC}_n = \underset{x_i \le m_n \le x_j}{\mathrm{med}} h(x_i, x_j) \tag{29}$$

where for all $x_i \neq x_j$, the kernel function h is defined as

$$h(x_i, x_j) = \frac{(x_j - m_n) - (m_n - x_i)}{x_j - x_i} \tag{30}$$

Given two values x_j and x_i, the difference between each of the values and the median is calculated. Then the difference between $(x_j - m_n)$ and $(m_n - x_i)$ is calculated, and the value is divided by $(x_j - x_i)$ to standardize it. Given the denominator$(x_j - x_i)$, MC_n will always lie between -1 and 1. Similar to the method of implementation for g_1, the medcouple of the bootstrapped samples becomes the bias-correction so that the confidence interval is equal to

$$\mathrm{BC}_{\mathrm{mc}} : (\hat{a}\hat{b}^{*(\alpha_{mc1})}, \hat{a}\hat{b}^{*(\alpha_{mc2})}) \tag{31}$$

where

$$\alpha_{mc1} = \Phi\left(2\hat{z}_{mc} + z^{(\alpha/2)}\right) \tag{32}$$

$$\alpha_{mc2} = \Phi\left(2\hat{z}_{mc} + z^{(1-\alpha/2)}\right) \tag{33}$$

The bias-correction is

$$\hat{z}_{mc} = MC_n \tag{34}$$

Transformations

Hall's transformation—Transformed normal approximation (T_1). Two version of Hall's transformation will be used—the normal approximation versus the bootstrapped version. Hall's transformation (1992) corrects for both bias and skewness and consists of transforming an asymmetric statistic into another statistic with a symmetric distribution. A one-sided $1-\alpha$ level confidence interval for $\hat{a}\hat{b}$ was originally presented by Hall (1992) as defined by

$$T_{1upp} = \left(-\infty,\ \hat{a}\hat{b} - n^{-1/2}\hat{\tau}g^{-1}(z_\alpha)\ \right) \tag{35}$$

$$T_{1low} = \left(\hat{a}\hat{b} - n^{-1/2}\hat{\tau}g^{-1}(z_{1-\alpha}), \infty\ \right) \tag{36}$$

where n is the sample size, $\hat{\tau}$ is the standard deviation,

$$g^{-1}(x) = n^{\frac{1}{2}}(a\hat{\gamma})^{-1}\left[\left(1 + 3a\hat{\gamma}\left(n^{-1/2}x - n^{-1}b\hat{\gamma}\right)\right)^{1/3} - 1\right] \tag{37}$$

and the z-score for the $100\alpha^{th}$ percentile value, $z_{1-\alpha}$ or z_α, is used in place of x in $g^{-1}(x)$. For the $g^{-1}(x)$ formula, $a = 1/3$, $b = 1/6$, and $\hat{\gamma}$ is the measure of skewness. The original measure for $\hat{\gamma}$ used by Hall (1992) is

$$\hat{\gamma} = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\left(X_i - \hat{\theta}\right)^3}{\hat{\tau}^3} \tag{38}$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\theta}\right)^2 \tag{39, 40}$$

where $\hat{\theta}$ is the mean and $\hat{\tau}^2$ is the variance. Without bootstrapping, however, $\hat{a}\hat{b}$ does not have a mean or variance. Therefore, the formulas for $\hat{\gamma}$, $\hat{\theta}$, and $\hat{\tau}^2$ cannot be used. Instead, a measure of skewness for the product of two normally distributed variables, as shown in Oliveira, Oliveira, and Seijas-Macias (2016), will be used so that $\hat{\gamma}$ is defined by

$$\hat{\gamma} = \frac{6\mu_a\mu_b\sigma_a^2\sigma_b^2}{\left(\mu_b^2\sigma_a^2 + \left(\mu_a^2 + \sigma_a^2\right)\sigma_b^2\right)^{3/2}}. \tag{41}$$

where μ_a and μ_b are the effects of a and b, respectively, and σ_a^2 and σ_b^2 are the standard errors of a and b. The standard error derived by Sobel (1982)

$$\sigma_{\hat{a}\hat{b}} = \sqrt{\hat{\sigma}_{\hat{a}}^2\hat{b}^2 + \hat{\sigma}_{\hat{b}}^2\hat{a}^2}. \tag{42}$$

will be used in place of equation (4) used to calculate $\hat{\tau}^2$. The confidence intervals for the original sample will be transformed using (35) and (36).

Bootstrap version of Hall's transformation method (T_2). Although Hall's transformation corrects for bias and skewness, the bootstrap version of Hall's transformation improves coverage accuracy. Zhou and Gao (2000) demonstrate how the bootstrapped version for constructing one-sided confidence intervals have better coverage for small to moderate samples (e.g., *n* = 12 or *n* = 100). The transformation will be applied to the bootstrapped sample to form two one-sided confidence intervals defined as

$$T_{2upp} = \left(-\infty,\ \hat{a}\hat{b} - n^{-1/2}\hat{\tau}^* g^{-1}\left(\hat{a}\hat{b}^{*(1-\alpha)}\right)\right) \tag{43}$$

$$T_{2low} = \left(\hat{a}\hat{b} - n^{-1/2}\hat{\tau}^* g^{-1}\left(\hat{a}\hat{b}^{*(\alpha)}\right), \infty\right) \tag{44}$$

where n is the sample size, $\hat{\tau}^*$ is the standard deviation of the bootstrapped sample,

$$g^{-1}(x) = n^{\frac{1}{2}}(a\hat{\gamma}^*)^{-1}\left[\left(1 + 3a\hat{\gamma}^*\left(n^{-1/2}x - n^{-1}b\hat{\gamma}^*\right)\right)^{1/3} - 1\right] \tag{45}$$

and the $100\alpha^{th}$ percentile value of the bootstrapped distribution, $\hat{a}\hat{b}^{*(1-\alpha)}$ or $\hat{a}\hat{b}^{*(\alpha)}$, is used in place of x in $g^{-1}(x)$. For the $g^{-1}(x)$ formula, $a = 1/3$, $b = 1/6$, and $\hat{\gamma}^*$ is the measure of skewness defined as

$$\hat{\gamma}^* = \frac{1}{n}\sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{\left(X_i - \hat{\theta}\right)^3}{\hat{\tau}^3} \tag{46}$$

$$\hat{\theta} = \hat{a}\hat{b}^*, \qquad \hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\theta}\right)^2 \tag{47, 48}$$

where $\hat{\theta}$ is the mean and $\hat{\tau}^2$ is the variance.

Transformed bootstrap-t (T_3). The bootstrap-t method is based on the *t* statistic rather than the estimate of $\hat{a}\hat{b}$. This resampling method, along with a transformed version called the bootstrap-Q, were evaluated in MacKinnon et al. (2004). The bootstrap-t uses the t statistic to form confidence intervals using

$$(\hat{\alpha}\hat{\beta} - T_{(1-\omega)/2} * \hat{\sigma}_{\hat{\alpha}\hat{\beta}},\ \hat{\alpha}\hat{\beta} - T_{\frac{\omega}{2}} * \hat{\sigma}_{\hat{\alpha}\hat{\beta}}) \quad (49)$$

where *T* is calculated by "dividing the difference between the bootstrap estimate and the original sample estimate by the bootstrap sample's standard error" (MacKinnon et al, 2004). The bootstrap-Q transforms the bootstrap-t by taking into account the skewness (s) of each bootstrap distribution of *T*, where

$$Q(T) = T + \frac{sT^2}{3} + \frac{s^2T^3}{27} + \frac{s}{6N} \quad (50)$$

And transforming the values back to *T* using

$$W(Q) = \frac{3\left[\left\{1+s[Q-\frac{s}{6N}]\right\}^{\frac{1}{3}}-1\right]}{s} \quad (51)$$

However, the confidence intervals formed by the bootstrap-t method tended to be too wide, while the bootstrap-Q had higher power and confidence intervals that were not as wide, but still had less power than the bias-corrected bootstrap.

The proposed alternative is Efron and Tibshirani's (1993) transformed bootstrap-t to fix the problem of a regular bootstrap-*t* interval from "performing erratically in small-sample, nonparametric settings" (p. 162). The procedure involves three nested levels of bootstrap sampling. The confidence interval for the transformed parameter is

$$\phi = .5\log(\frac{1+\theta}{1-\theta}) \quad (52)$$

followed by an inverse transformation of the endpoints with

$$\frac{e^{2\phi}-1}{e^{2\phi}+1}\,. \quad (53)$$

While bootstrap-*t* procedures have confidence intervals that are often too wide (Efron & Tibshirani, 1993; MacKinnon, et al., 2004), the transformed percentile-*t* bootstrap forms much shorter intervals. In comparison with the bootstrap-Q, instead of taking skewness into account, this alternative transformation corrects for the endpoints. Additionally, the performance of the transformed percentile-*t* bootstrap is more stable with small sample sizes.

For the first-level bootstrap, $B_1 = 1000$ bootstrap samples will be generated to estimate $\hat{a}\hat{b}$. There will be $B_2 = 25$ second-level bootstrap samples, as suggested by Efron and Tibshirani (1993), to estimate the standard error. The distribution of the bootstrap-*t* distribution will be estimated using $B_3 = 1000$ new bootstrap samples. Finally, the endpoints of the interval will be mapped back to the original $\hat{a}\hat{b}$ scale using (51).

Box-Cox transformation (T_4). The Box-Cox transformation was a proposed modification by Box and Cox (1964) to Tukey's (1957) family of power transformations. The Box-Cox transformation is a way to transform non-normal data into normal data, allowing assumptions for normal data to be used. Osborne (2010) also notes that the Box-Cox transformation both normalizes skewed data and improves effect sizes. The method eliminates the need for the researcher to blindly test several different transformations for the best one by providing a family of transformations to work with.

Applying the transformation involves first anchoring the minimum value of the distribution to 1. This is followed by calculating the value of λ, the Box-Cox transformation coefficient, for which to raise the variables to determine optimal reduction of skewness. Osborne (2010) provides guidelines on how to calculate λ by hand. There are statistical packages, however, that have Box-Cox implemented and iteratively estimate the best λ options. After λ is estimated, the following equations are used to transform the data

$$y_i^{(\lambda)} = \begin{cases} \frac{\{(y_i+\lambda_2)^{\lambda_1}-1\}}{\lambda_1}; & \lambda_1 \neq 0 \\ \log(y_i+\lambda_2); & \lambda_1 = 0 \end{cases} \tag{54}$$

where "λ_1 is the transformation parameter and λ_2 is chosen such that y_i is greater than $-\lambda_2$. Although this transformation takes into account the discontinuity at $\lambda = 0$, it does not account for negative observations. Manly (1976) proposed the following alternative that will be used to transform the bootstrapped data

$$y_i^{(\lambda)} = \begin{cases} (\exp(\lambda y_i)-1)/\lambda\ ; & \lambda \neq 0 \\ y_i; & \lambda = 0 \end{cases}. \tag{55}$$

Since the transformed data is expected to be nearly symmetrical and normally distributed, there are two methods to test for calculating confidence intervals.

Box-Cox transformation—percentile method (**T_{4p}*****).*** One way to calculate the confidence interval for the transformed data is to assume the transformed data are normally distributed and use the percentile method. The confidence intervals will be calculated by obtaining values at the corresponding $100\alpha th$ percentiles. Using $\alpha =$ .05, the confidence interval is

$$\hat{a}\hat{b}_{T4}^{*(.025)} \leq \hat{a}\hat{b}_{T4} \leq \hat{a}\hat{b}_{T4}^{*(.975)} \tag{56}$$

where $\hat{a}\hat{b}_{T4}^{*(.025)}$ and $\hat{a}\hat{b}_{T4}^{*(.975)}$ are the 25th and 975th values on the bootstrapped transformed distribution.

Box-Cox transformation—z-score method (**T_{4z}*****).*** Another method to calculate confidence intervals is to assume the transformed data have been transformed into a standard normal distribution. The confidence intervals will be calculated using

$$CI: \hat{a}\hat{b}_{T4} \pm z_{(1-\alpha/2)} * \frac{s_{T4}}{\sqrt{n}} \quad (57)$$

where $\hat{a}\hat{b}_{T4}$ is the Box-Cox transformed value of $\hat{a}\hat{b}$, $z_{(1-\alpha/2)}$ is the z critical value on a standard normal distribution with the specified α Type I error rate, s_{T4} is the sample standard deviation from the transformed bootstrap samples, and n is the number of bootstrap replications.

Research Questions

1. Which bias-correction alternative(s) best improve(s) accuracy of the Type I error rate, ideally .05, compared to Efron & Tibshirani's original bias-corrected bootstrap method?
    - For the alternative measures of bias, it is hypothesized that the medcouple will have the most accurate Type I error rates compared to the median, mean, and g_1 because it is bounded between -1 and 1. The bounds for the medcouple will limit the value of the bias-correction, potentially keeping it smaller than g_1, which is unbounded. For the transformations, it is hypothesized that all of the alternatives will have more accurate Type I error rates compared to the original bias-corrected bootstrap because each have been shown to produce narrower confidence intervals.
2. How will sample size for each of the bias-correction alternatives affect Type I error rates compared to the original z_0 bias-corrected bootstrap method?
    - It is hypothesized that for the alternative measures of bias, the mean and g_1 will be most affected by sample size, as their calculations involve sample size; there may be more variability in Type I error rates for these alternatives depending on sample size. Hall's transformation, the bootstrap

version of Hall's transformation, and the transformed bootstrap-*t* have all been shown to have more consistent coverage in small sample sizes. Therefore, it is hypothesized that these three alternatives will produce lower Type I error rates in smaller sample sizes compared to the Box-Cox transformation.

3. Which combinations of sample size and effect size for each of the bias-correction alternatives will produce more accurate Type I error rates?
    - It is hypothesized that for the alternative measures of bias, g_1 and the medcouple will produce more accurate Type I error rate than the mean and median in smaller sample sizes for the medium and large effect sizes. As observed in previous studies, the alternative measures of bias are hypothesized to produce more erratic measures of Type I error rate for a small effect size, such that there is a smaller Type I error rate with a small sample size which increases to become inflated at larger sample sizes. It is hypothesized that for the transformations, the alternatives will produce decreased Type I error rate for smaller sample sizes in medium and large effect sizes. It is also hypothesized that the alternatives will produce increased Type I error rates for smaller sample sizes for a small effect size.
4. How will power be affected if the Type I error rate is found to remain constant at $\alpha = .05$ for each of the alternative measures of bias?
    - It is hypothesized that power will be negatively affected (power will decrease as compared to the percentile bootstrap and the joint significance test as baseline measures) for the alternative measures of bias compared to

the transformation alternatives because the transformation alternatives provide more control over the actual range of the confidence intervals.

Methods

The simulation consists of two parts: data generation and application of alternatives.

Data Generation

To determine whether the proposed alternatives will maintain more accurate Type I error rates compared to the bias-corrected and accelerated bias-corrected bootstrap methods, a simulation will be performed using R (R Core Team, 2016). This study is an extension of Fritz et al. (2012), so the first four factors that were varied in the previous study will also be varied in this study to focus on the effects of the bias adjustments and alternatives.

The first factor that will be varied is the test of mediation; the original bias-correction using z_0 will be used to test for mediation in the data. The percentile bootstrap and joint significance test will also be used to replicate results of previous studies (MacKinnon et al., 2004) and used as control factors. Each of the eight alternative measures of bias will then be integrated and tested for significance, for a total of 11 different tests of mediation.

The second and third factors to be varied are the path effect sizes of *a* and *b*. Using Cohen's (1988) guidelines for small, medium, and large effect sizes, *a* and *b* will alternately be set to 0, 0.14, 0.39, or 0.59, forming 16 different effect size combinations.

The fourth varied factor will be sample size, selected to represent the range of commonly used sample sizes in the social sciences: 50, 100, 500, and 1000. Fritz et al. (2012) considered an additional sample size of 2500 but found that as sample sizes approached 2500, Type I error rate returned to .05 for all tests including the bias-corrected bootstrap using z_0.

The fifth factor that Fritz et al. (2012) examined will not be varied in this study due to the finding that the number of bootstrap samples does not affect Type I error rates. The number of bootstrap samples will be set at 1000.

The RNORM function in R will be used to generate 1000 values of *X*. Values for *M* and *Y* will be generated using the set values for a and b (as variations of the second and third factors) through regression equations 2 and 3. Residuals will also be generated using the RNORM function for *M* and *Y*. Each original sample will be generated according to the parameters set by each of the varied factors described above, and the confidence interval for that sample will be calculated. Additionally, the bias size for the proportion of observations below the true mediated effect will not be controlled for. Therefore, the actual size of bias for the simulated samples will be saved as an additional outcome variable.

The original sample will be bootstrapped 1000 times, each of the methods will be applied, and the confidence intervals will be calculated for each bootstrap sample. The process will be repeated 1000 times, generating 1000 replications.

Outcome Variables

The rejection rate is the number of times zero is outside the confidence intervals. Rejection rate is coded as '0' when zero is within the confidence interval and '1' when zero is outside the confidence interval. The rejection rate is *Type I error* when the population effect $ab = 0$ and *power* when $ab \neq 0$. *Coverage* is the number of times the true mediated effect falls within the confidence intervals. Coverage is coded as '0' when *ab* falls outside the confidence interval and ‘1’ when *ab* falls inside the confidence interval. *Balance* assesses how many times the true mediation effect falls to the left or the right of the confidence interval. Balance is coded as ‘-1’ if *ab* falls to the left, determined by the number of times the lower bound of the confidence interval is greater than the observed effect. Balance is coded as ‘1’ if *ab* falls to the right, when the upper bound of the confidence interval is less than the observed effect. Bias is the proportion of bootstrap replications of $\hat{a}\hat{b}$ that are less than the original estimate of the mediated effect $\hat{a}\hat{b}$.

Data Analysis

The data will be analyzed using PROC GLIMMIX in the SAS® 9.4 software (SAS Institute, 2012). Generalized linear mixed models (GLMM) will be used to fit the data for Type I error rate and power due to the binomial nature of the outcome variables and the presence of both fixed and random effects. For Type I error and power, the effects of *a* and *b* will be tested separately. Since either *a* or *b* is required to equal zero for each condition, collapsing across one effect size will produce misleading marginal means (Fritz et al., 2012). Type I error rate and power for each of the alternatives will be compared to the control methods to determine significant differences and moderation effects. Type I error rate and power for type of test by sample size by effect size interactions will also be tested. Finally, the effect of bias and sample size on Type I error and power will be tested. Scheffé correction will be used as a Type I error correction due to the exploratory nature of the contrasts.

Results

Type I Error

Type I error was analyzed using a method × sample size × *a* GLMM. There was a significant three-way interaction, $F(117, 156) = 3.92$, $p < .01$. Similarly, when Type I error was analyzed using method × sample size × *b*, there was a significant three-way interaction, $F(117, 156) = 4.70$, $p < .01$. This suggests that there is a difference in interactions among methods. The main effect of method on Type I error rate, collapsed across sample size and effect size, is presented in Figure 3. The figure shows that the control methods (joint significance test, percentile bootstrap, and z_0), as well as T_{4p}, z_{mc}, z_{mean}, and z_{medi} have Type I error rates far below the other methods. The main of effects of only the methods with Type I error rates closer to the targeted .05 are presented in Figure 4 for a closer visual representation on their overall pattern.

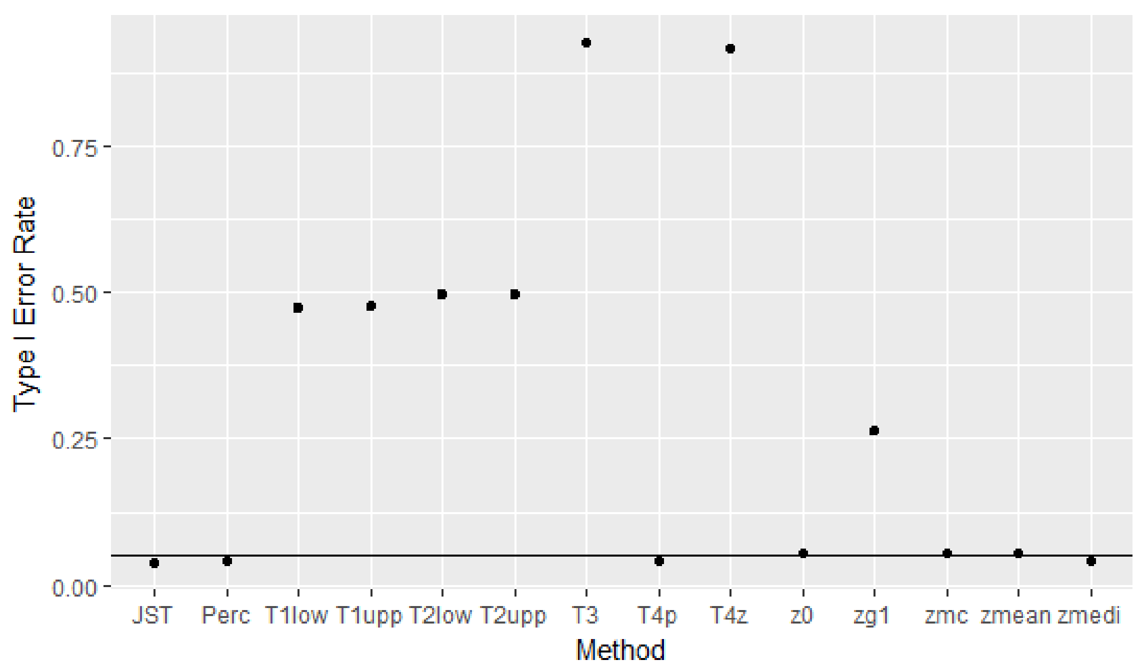

Figure 3. Main effect of method on Type I error rate collapsed across sample size and effect size.

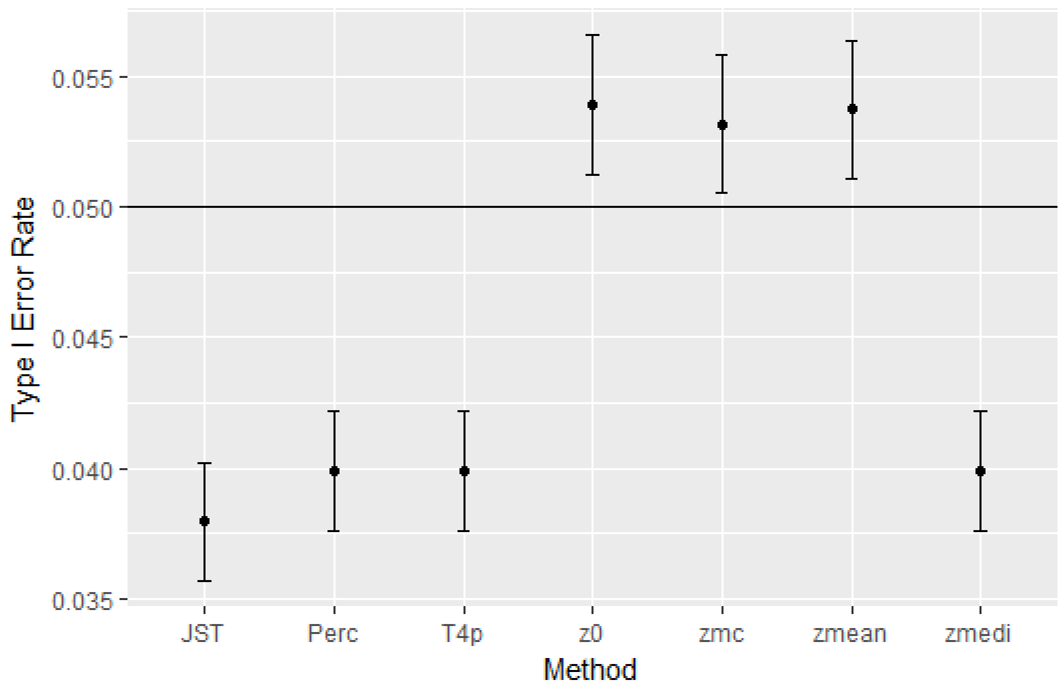

Figure 4. Main effect of method on Type I error rate collapsed across sample size and effect size using only the remaining methods.

The goal of this study was to find alternative bias corrections to the bias-corrected bootstrap with better accuracy in Type I error. Therefore, using the descriptive statistics for Type I error rates (reported in Table 1), there were no further contrasts conducted on T_{1low}, T_{1upp}, T_{2low}, T_{2upp}, T_3, T_{4z}. These eliminated methods had a minimum Type I error rate above the targeted .05 rate, the smallest minimum was 0.436.

Table A.1 reports the F values associated with the method × sample size × *a* three-way interaction effect sliced by method. Table A.2 reports the F values associated with the method × sample size × *b* three-way interaction effect sliced by method. The F values for the remaining alternative methods ($z_{mean}, z_{medi}, z_{g1}, z_{mc}, and\ T_{4p}$) were all statistically significant, suggesting that within each method, the effect of sample size is dependent on the effect size of *a* or *b*, respectively.

Table 1

Descriptive statistics for Type I error for each of the methods

Method	*M*	*SD*	*Min*	*Max*
Joint significance test	0.0380	0.0218	0.001	0.072
Percentile	0.0399	0.0228	0.000	0.072
z_0 (bias-corrected)	0.0539	0.0288	0.004	0.106
z_{mean}	0.0537	0.0289	0.005	0.106
z_{medi}	0.0399	0.0228	0.000	0.072
z_{g1}	0.2639	0.1626	0.051	0.522
z_{mc}	0.0531	0.0281	0.005	0.103
T_{1low}	0.4744	0.0161	0.436	0.500
T_{1upp}	0.4765	0.0149	0.455	0.508
T_{2low}	0.4975	0.0120	0.474	0.521
T_{2upp}	0.4986	0.0130	0.477	0.524
T_3	0.9269	0.0398	0.801	0.962
T_{4p}	0.0399	0.0228	0.000	0.072
T_{4z}	0.9186	0.0530	0.801	0.967

Figure 5 displays the main effect of sample size collapsed across all methods and sample sizes. Although there is a decreasing trend in Type I error as sample size increases, there were no statistically significant differences in Type I error rate among sample sizes. Figure 6 displays the main effect of effect size collapsed across all methods and sample size. Similarly, there were no statistically significant differences in Type I error rate among effect sizes.

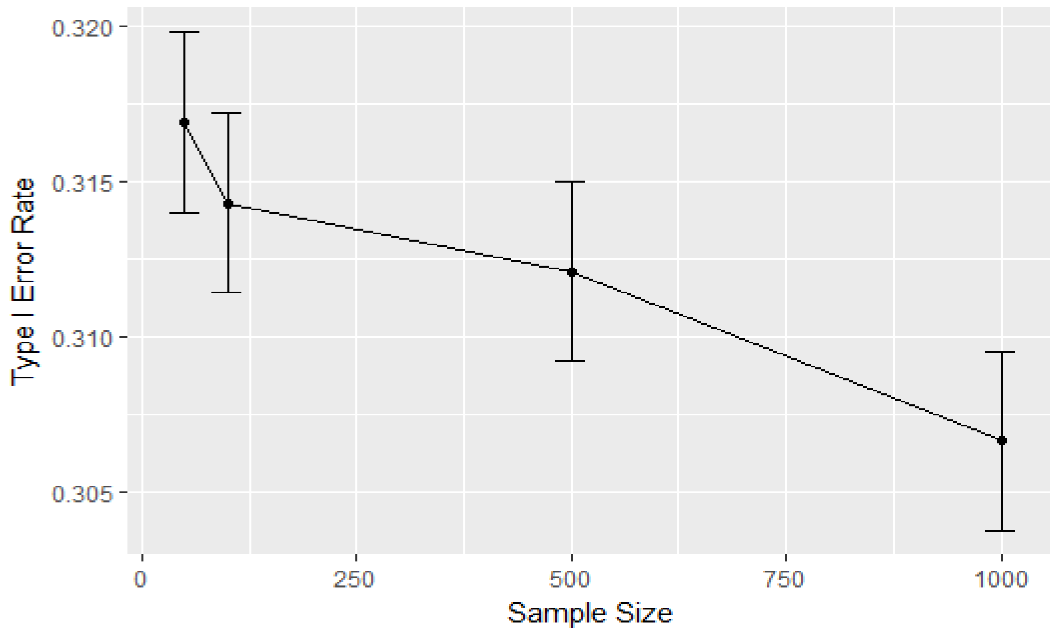

Figure 5. Main effect of sample size collapsed across method and effect size.

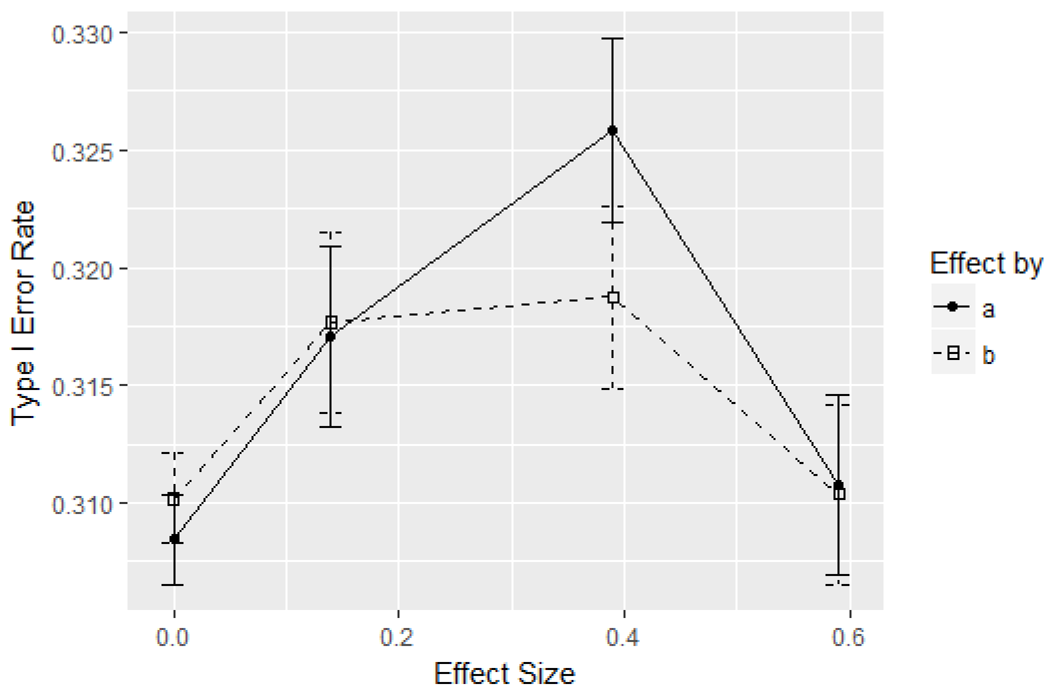

Figure 6. Main effect of effect size collapsed across method and sample size.

For Figures 7.1 thru 9.2, two-way interaction plots for method $\times$ a, spliced by sample size, were created to illustrate the general patterns of the remaining alternatives. Only conditions where $b = 0$ were presented in these figures. Similar patterns were observed when examining two-way interaction plots for method $\times$ b when $a = 0$.

Figure 7.1 shows each of the remaining alternative measures of bias methods compared to the joint significance test, Figure 8.1 compares the methods to the percentile bootstrap, and Figure 9.1 compares the methods to the original bias-corrected bootstrap. In each of these three figures, visual inspection revealed that z_{g1} has Type I error rates higher than the other alternatives, with the exception at 0.39 and 0.59 effect size at a sample size of 1000. Figures 7.2, 8.2, and 9.2 display results without z_{g1} for a closer visual on the lower Type I error rates.

Visual inspection of Figure 9.2 in conjunction with examining descriptive statistics showed that z_{med} and T_{4p} consistently had lower Type I error rates compared to the other methods, particularly with medium and large effect sizes with 50 and 100 sample sizes. Contrasts for the median correction and the T_{4p} correction compared to the bias-corrected bootstrap were tested to determine whether the median had significantly lower Type I error rates. The differences between the median correction and the original bias-corrected bootstrap at all sample sizes were not significant. The differences between z_{medi} and T_{4p} compared to z_0 at all levels of a were also not significant. Additionally, Type I error rates for z_{medi} and T_{4p} were the same as the percentile bootstrap rates.

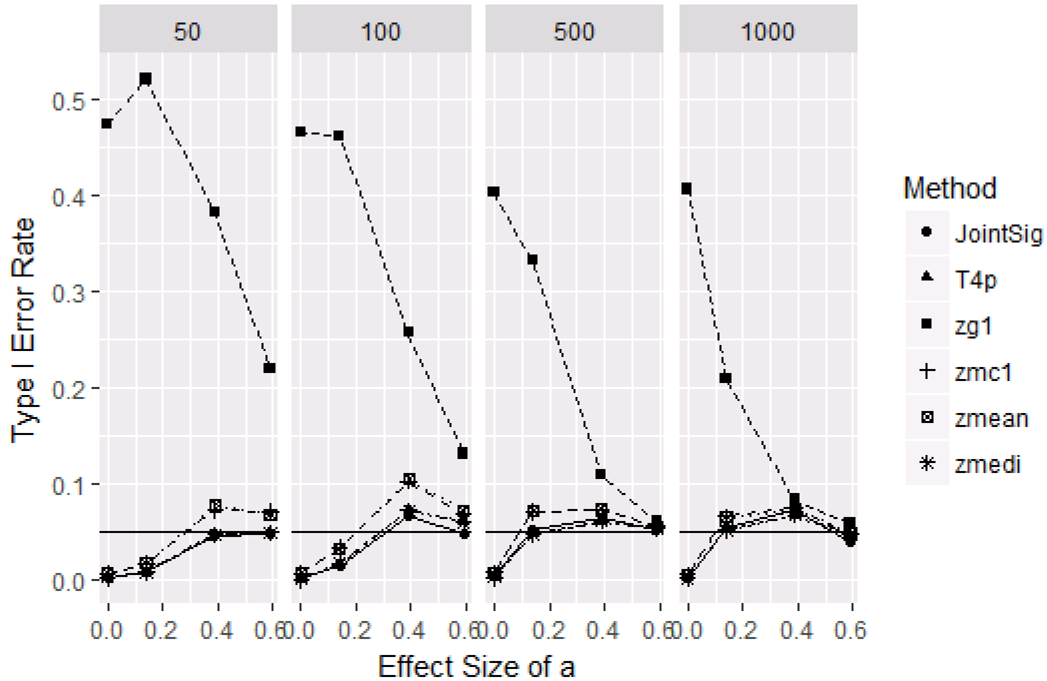

Figure 7.1. Alternative measures versus joint significance test (JointSig). Comparison of alternative measures to the joint significance test for effect size of a and the Type I error rate, spliced by sample size.

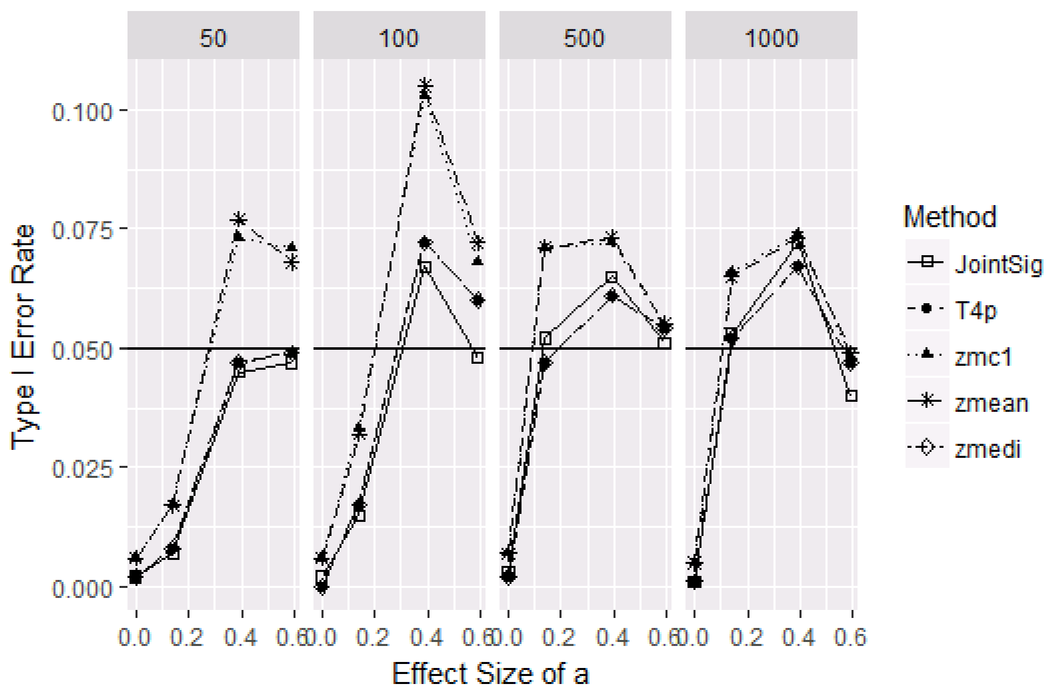

Figure 7.2. Alternative measures versus joint significance test (without z_{g1}) for effect size of a and the Type I error rate, spliced by sample size.

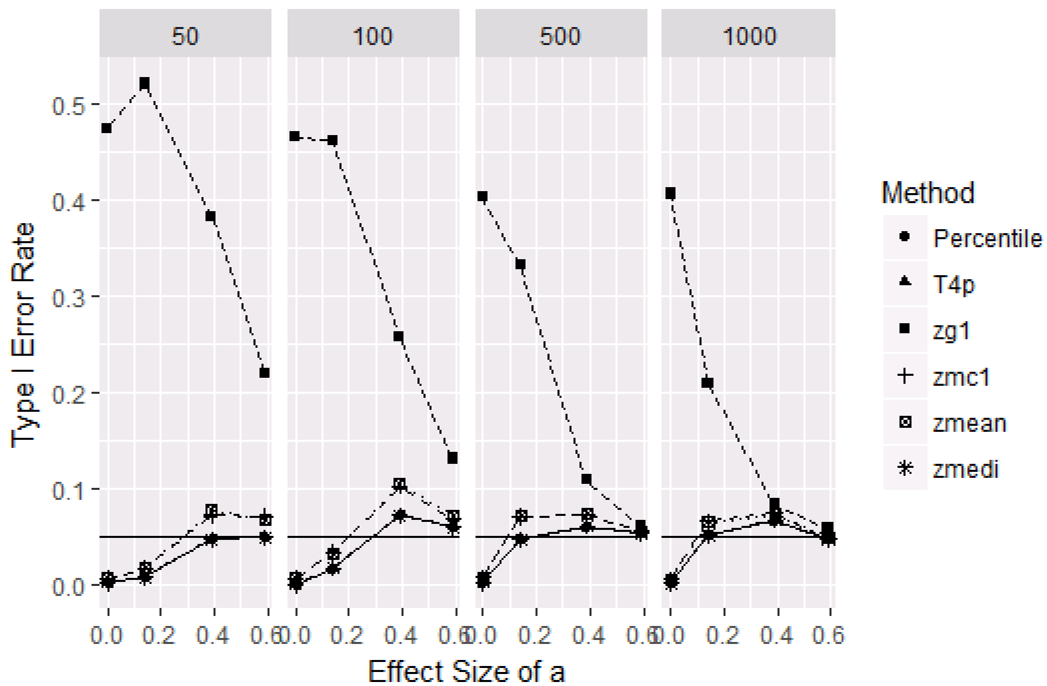

Figure 8.1. Alternative measures versus percentile bootstrap. Comparison of alternative measures to the percentile bootstrap for effect size of a and the Type I error rate, spliced by sample size.

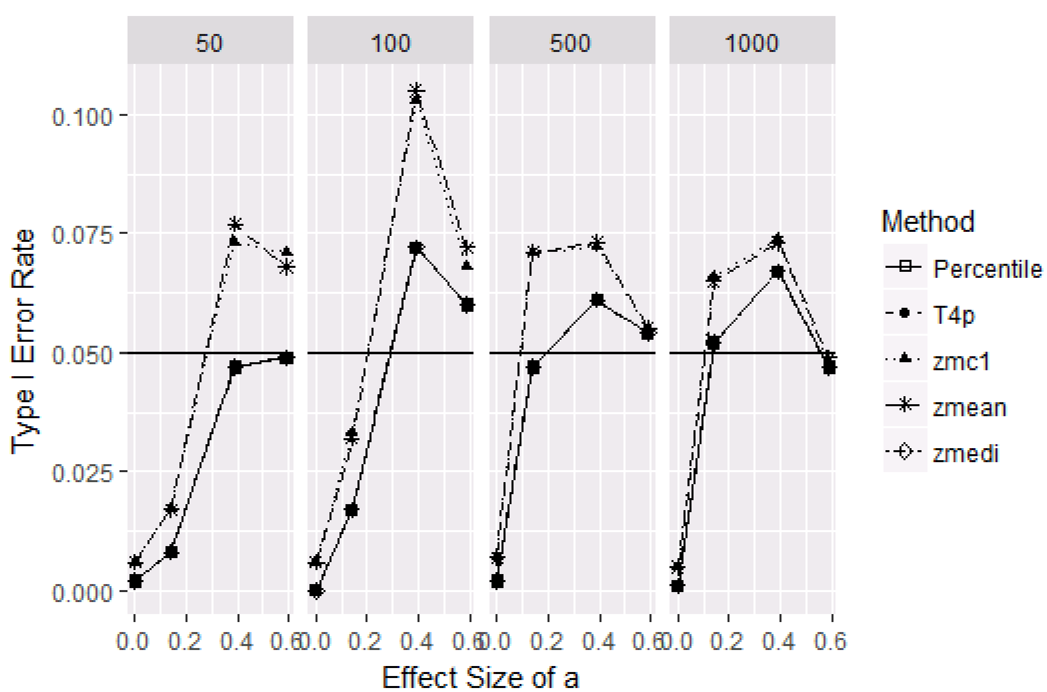

Figure 8.2. Alternative measures versus percentile bootstrap (without z_{g1}) for effect size of a and the Type I error rate, spliced by sample size.

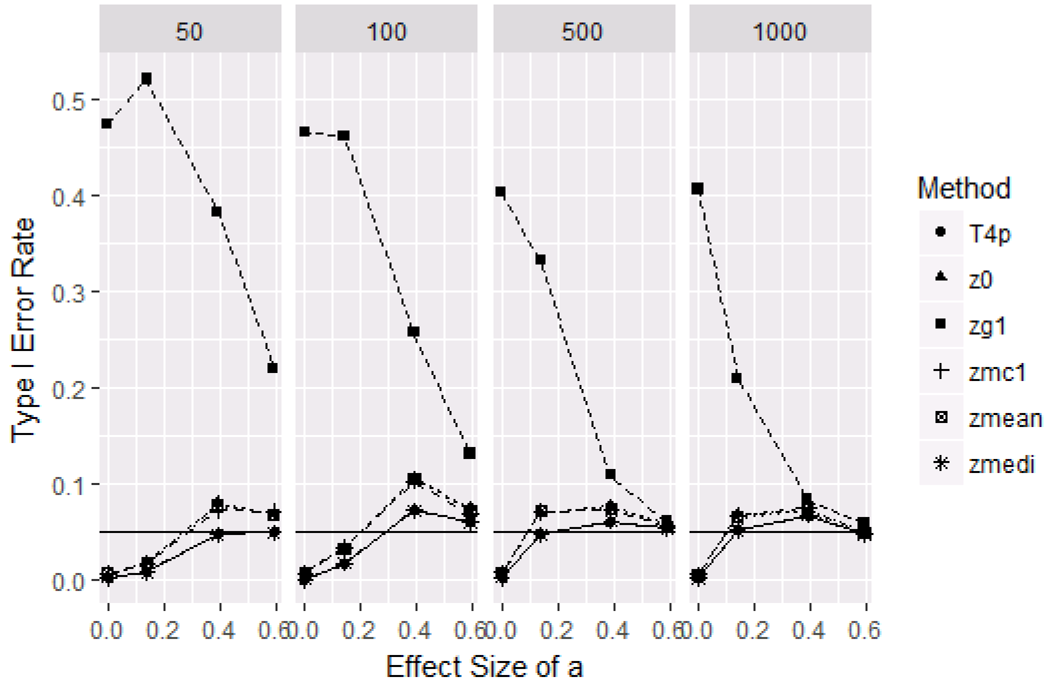

Figure 9.1. Alternative measures versus bias-corrected bootstrap. Comparison of alternative measures to the bias-corrected bootstrap (z_0) for effect size of a and the Type I error rate, spliced by sample size.

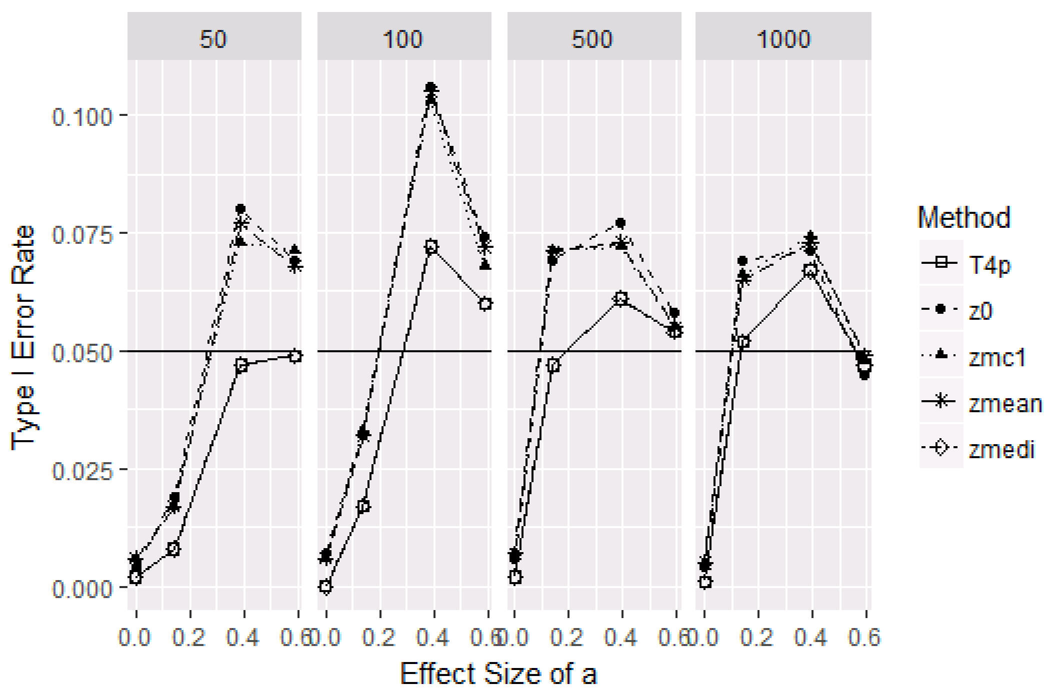

Figure 9.2. Alternative measures versus bias-corrected bootstrap (without z_{g1}) for effect size of a and the Type I error rate, spliced by sample size.

Power

Power was analyzed using a method × sample size × a GLMM using only the methods considered in the Type I error rate section. There were two significant interactions: method × sample size, $F(21, 168) = 22.35$, $p < .01$, and method × a, $F(14, 168) = 6.75$, $p < .01$. When power was analyzed using method × sample size × b, again keeping only eight methods as predictors, the model did not converge. Instead, the three-way interaction was taken out as a predictor. Thus, three main effects were tested (method, sample size, and b) and three interactions were tested (method × sample size, method × b, and sample size × b). There were two significant interactions: method × sample size, $F(21, 210) = 30.32$, $p < .01$, and method × b, $F(14, 210) = 20.93$, $p < .01$.

Power for the remaining methods × a interactions are displayed in Figure 10. Each column is a different effect size b and each row is a different sample size. There is an increasing trend in power by sample size; as the sample size increases, power also tends to increase. There is also an increasing trend in power by b; as b increases, power also tends to increases.

Contrasts were analyzed for the following methods compared to the control conditions: $z_{mean}, z_{medi}, z_{mc}, and\ T_{4p}$. There were no significant differences in power between the four conditions compared to the control conditions.

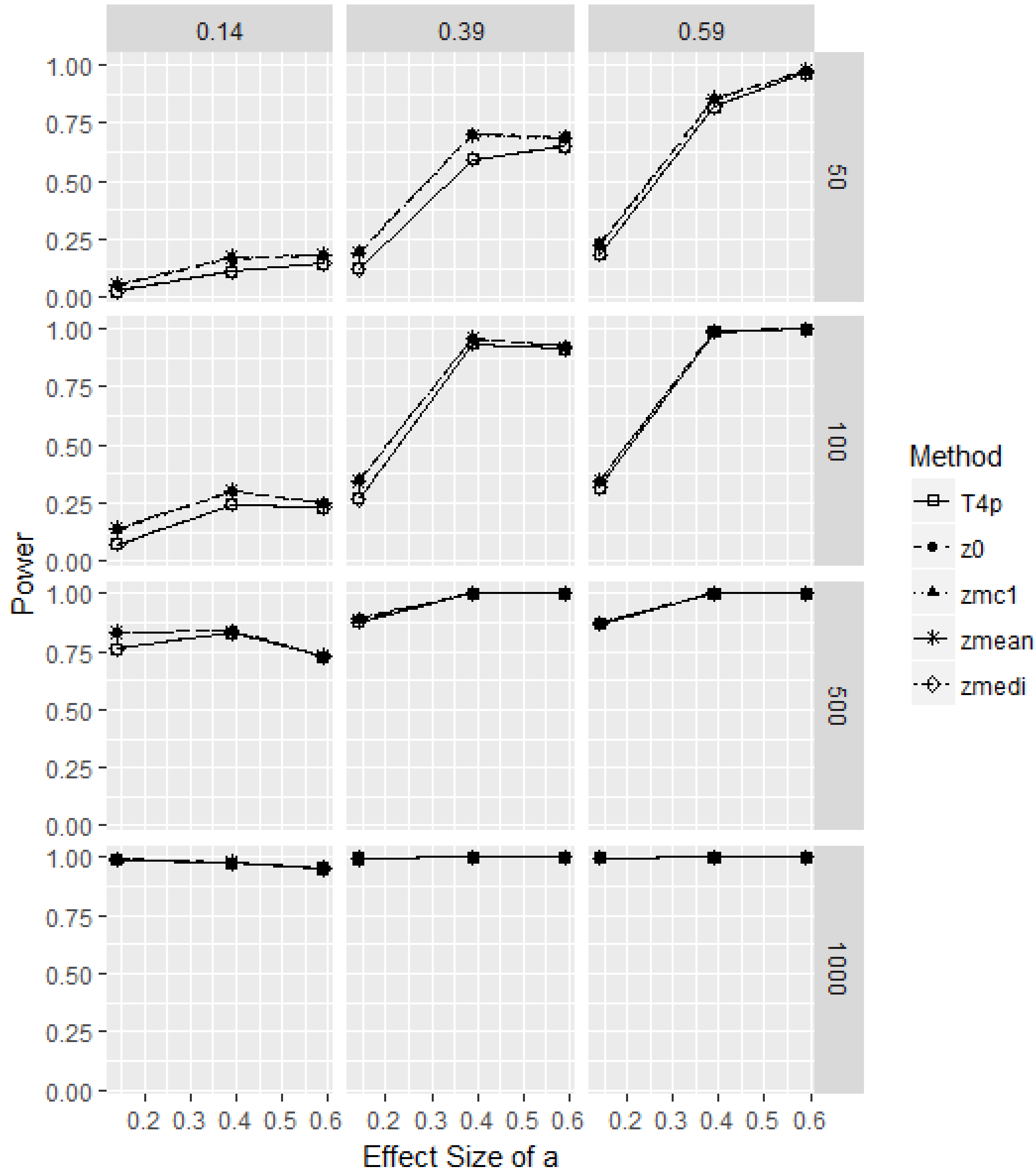

Figure 10. Power for the bias-corrected bootstrap by method for sample size of 50 for each small, medium, and large effect size of a, spliced by small, medium, and large effect size of b.

Coverage

The coverage for the alternative methods are reported in Table A.4. There is a wide range of coverage between all of the methods. The percentile bootstrap, z_{medi}, and T_{4p} all had the highest coverage rate of 95.14%. z_0, z_{mc}, and z_{mean} had the next highest coverage rates, with a range from 94.25% to 94.32%. T_3 had the lowest coverage of 6.08% followed by T_{4z} with 6.20%.

Balance

The percentage of times the confidence interval failed to the left or the right—balance—is reported in Table A.5. With the exception of T_{1upp}, T_{2upp}, and T_3, the left failure percentage was higher than the right failure percentage.

Bias

The results for the effect of bias (*M* = 0.03, *SD* = 0.10, *Minimum* = -0.59, *Maximum* = 0.63) on Type I error, using only the eight remaining alternatives was not found to be significant. There was also no significant effect of bias on power.

Discussion

The results of this study were similar to findings by MacKinnon et al. (2004), Cheung (2007), and Fritz et al. (2012) that the bias-corrected bootstrap tests of mediation have elevated Type I error rates in conditions where the sample size is less than 500 with medium or large effect sizes of a. The results are discussed in order of the original research questions.

1. Which alternative measure(s) of bias best improve(s) accuracy of the Type I error rate, ideally .05, compared to the original bias-corrected bootstrap method?

For the alternative measures of bias, it was hypothesized that the medcouple would have the most accurate Type I error rates compared to the median, mean, and g_1 because it is bounded between -1 and 1. The medcouple, however did not have the most accurate Type I error rate. Instead, the medcouple shared similar patterns in Type I error rate elevation compared to the original bias-corrected bootstrap and the mean measure. These similar patterns may have been observed because of the similar method z_0, the mean, and the medcouple were implemented. The z_0 used the proportion under the mediated effect, while the mean and medcouple were variations of the mediated effect. Thus, the difference in bias-correction was not enough to decrease Type I error rate.

For the transformations, it was hypothesized that all of the alternatives will have more accurate Type I error rates compared to the original bias-corrected bootstrap because each have been shown to produce narrower confidence intervals. This was not the case with the study results; the transformations produced even higher levels of Type I error rate. In the case of transformed bootstrap-*t* (T3), the confidence intervals became too narrow, rendering it the method with the greatest Type I error.

2. How will sample size for each of the alternative measures of bias affect Type I error rates compared to the original bias-corrected bootstrap method?

It was hypothesized that for the alternative measures of bias, the mean and g_1 would be most affected by sample size, as their calculations involve sample size. Similar results to Fritz et al. (2012) for the control conditions, $z_{mean}, z_{medi},\ z_{mc}, and\ T_{4p}$ were found in that Type I error rates were elevated for a medium or large effect size of *a* for *n* = 50, 100, 500, and 1000. The transformation alternatives were not further examined due to their inflated Type I error rates.

3. For which combinations of sample size and effect size can each of the alternative measures of bias correct for Type I error rate?

No alternative was found to significantly decrease Type I error rate in comparison to the original bias-corrected bootstrap.

4. How will power be affected if the Type I error rate is found to remain constant at $\alpha =$ .05 for each of the alternative measures of bias?

Power is closely tied to Type I error rate, sample size, and effect size. Results similar to findings by Fritz et al. (2012) were observed for the control conditions, $z_{mean}, z_{medi}, z_{mc}, and\ T_{4p}$. Serlin (2000) suggests a range of .04 - .06 Type I error rate for a method to be considered robust. The transformation methods had Type I error rates that were well outside of the suggested range. Although the remaining methods fell within the range for robustness, differences in power were not further explore due to the goal of this study to find alternatives for decreasing Type I error.

An interesting finding was that the Box Cox transformation with the percentile bootstrap method (T_{4p}) produced identical Type I error rates and power when compared to the percentile bootstrap method. The actual confidence intervals, however, differed slightly. This suggests that though the data were transformed to correct for skewness, perhaps the problem surrounding the issue of the elevated Type I error rates lies beyond merely finding a better measure of bias.

Empirical Example

To illustrate the similarities, differences, and patterns found in the alternative measures of bias, the alternatives examined in this study will be applied to data from the Athletes Training and Learning to Avoid Steroids (ATLAS) program (Goldberg et al., 1996). The ATLAS program presents two healthy alternatives to high school football players: healthy nutrition behaviors and appropriate strength training, as alternatives to anabolic steroid use. Data were collected from players on different measures including intentions to use anabolic steroids, nutrition behaviors, and strength training self-efficacy at three time points (start of the football season, end of football season, and one year follow-up). MacKinnon et al. (2001) examined possible mediators from the ATLAS program, of which, the relation between participation in the ATLAS program (X) and a student's intentions to use anabolic steroids measured 9 months after finishing the ATLAS program (Y) is mediated by a student's perceived susceptibility to the adverse effects of steroid use immediately after completing the ATLAS program (M).

After deleting cases with missing data, a complete sample of 731 students were used for this analysis. Cases with missing data were deleted because of the nature of bootstrapping. Since the same case could potentially be selected more than once, there is the possibility of having an entire bootstrap sample of missing data. Table 2 contains the confidence intervals formed by each method. The estimated value $\hat{a} = 0.5949$, $\hat{b} = -0.0961$, $\hat{a}\hat{b} = -0.0572$. As expected, the percentile and the median produce identical confidence intervals. The original bias-corrected bootstrap, mean, and medcouple produced similar results. Additionally, the Box-Cox percentile produced confidence intervals similar to the percentile but not identical.

Table 2

Confidence intervals for the ATLAS data for each method

Method	*Lower CI*	*Upper CI*
Percentile	-0.1044	-0.0239
z_0 (bias-corrected)	-0.1085	-0.0250
z_{mean}	-0.1086	-0.0251
z_{medi}	-0.1044	-0.0239
z_{g1}	-0.1290	-0.0373
z_{mc}	-0.1069	-0.0244
T_{1low}	-0.0579	∞
T_{1upp}	$-\infty$	-0.0564
T_{2low}	-0.0572	∞
T_{2upp}	$-\infty$	-0.0572
T_3	-0.0590	-0.0556
T_{4p}	-0.0942	-0.0233
T_{4z}	-0.0552	-0.0529

Conclusion

The purpose of this study was to consider alternative bias-corrections in the bias-corrected bootstrap to reduce Type I error rates in the elevated conditions without reducing power. None of the tested alternative measures of bias, however, could produce more accurate Type I error rates in the elevated conditions.

Limitations

One limitation of this study was not controlling for the bias size of bootstrapped observations falling under the true mediated effect. The bias size was free to vary and the descriptive statistics were reported for the outcome variable. There were no guidelines for interpreting the size of bias. Negative versus positive bias effects were also unexamined.

A second limitation is concerning the transformed bootstrap-*t* (T3). The second and third nested bootstrap level sizes were based off the recommendation of Efron and Tibshirani (1993). The recommendations, however, were not specific to mediation analysis and therefore did not pertain directly to the distribution of a product. Without further study into the effect of the nested bootstrap level sizes, it can only be speculated that the chosen level sizes contributed to the extreme narrowness of the newly formed confidence intervals.

A third limitation is that alternative methods for implementing the bias-correction were not examined or applied. The alternative measures of bias for this study were implemented in ways similar to the original BC bootstrap where the proportion of observations below a specified value were used to calculate a corresponding z-score. Perhaps a different method of implementing a bias-correction altogether can be tested in future studies to account for size and direction of skew of the distribution.

A fourth limitation is that the methods in this study consisted only of measures of skew and transformations. Additionally, these measures of skew and transformations were not originally developed to be used in a statistical mediation analysis context. The initial intent of many of these methods was to correct for skewed distributions of single variables, whereas the major focus of this study was on correcting for the skewed distribution for the product between two variables. One alternate approach would be to focus on how *M* being both a dependent and causal variable affects Type I error and power.

Future Directions

This study could only cover a handful of alternatives to reducing Type I error rate, yet there are more ways of approaching this issue to consider. Drawing from the limitations, the following future directions are delineated. A closer study of bias size and effect on Type I error can be conducted. Although descriptive statistics were run for the measure of bias in each of the bootstrap samples, work can be dedicated towards defining the bounds between bias sizes and what constitutes a negligible, small, medium, or large bias size before examining the effect of bias size on Type I error rate, power, and effect size.

Another possible direction for this study is to consider different approaches in alternative measures of bias. One such method is the iterated prepivoted bootstrap that combines the two different methods tested in this study. The first prepivot removes estimated bias, and with each prepivoted iteration, higher order corrections are automatically made, leading to smaller coverage error (Beran, 1987). Future studies can be conducted to test the effects of combing a bias-correction alternative with transforming data to be normally distributed. The bias-correction alternatives are not limited to different measures of z_0, nor are they limited to transformations. A more comprehensive study on different correction methods can inform future studies testing different alternatives.

Based on these findings, more work needs to be done to find a better measure of bias for the bias-corrected bootstrap test of mediation. Alternatively, a completely different approach may be necessary to examine the anomalous findings of the bias-corrected bootstrap test of mediation. Until a better method is found, however, in situations where the bias-corrected bootstrap is utilized for the increased power, researchers should also be wary of the increased Type I error rate and the potential implications this finding could have on their own study's applications.

References

Baron, R.M., and Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika, 74,* 457-468.

Box, G.E.P., & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26*(2), 211-252.

Brys, G., Hubert, M., & Struyf, A. (2003). A comparison of some new measures of skewness. In R. Dutter, P. Filzmoser, U. Gather, & P.J. Rousseeuw (Eds.), *Developments in Robust Statistics.* Paper presented at the International Conference on Robust Statistics 2001 (ICORS 2001), Vorau, Austria (pp. 98-113). Springer-Verlag Berlin: Heidelberg GmbH.

Brys, G., Hubert, M., & Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics, 13*(4), 996-1017.

Cheung, M. W.-L. (2007). Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling, 14,* 227-246.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*, 1-26.

Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York, NY: Chapman & Hall.

Fritz, M.S., & MacKinnon, D.P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18,* 233-239.

Fritz, M.S., Taylor, A.B., & MacKinnon, D.P. (2012). Exploration of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research, 47,* 61-87.

Goldberg, L., Elliot, D., Clarke, G.N., MacKinnon, D.P., Moe, E., Zoref, L., … Lapin, A. (1996). Effects of a multidimensional anabolic steroid prevention intervention: The Adolescents Training and Learning to Avoid Steroids (ATLAS) program. *Journal of the American Medical Association, 276,* 1555-1562.

Hall, P. (1992). On the removal of skewness by transformation. *Journal of the Royal Statistical Society*, *54*(1), 221-228.

Higgins, J.J. (2004). *Introduction to Modern Nonparametric Statistics.* Belmont, CA: Brooks/Cole Cengage Learning.

Joanes, D.N., & Gill, C.A. (1998). Comparing measures of sample skewness and kurtosis. *The Statistician, 47,* 183-189.

Judd, C.M., & Kenny, D.A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*(5), 602-619.

Lomnicki, Z.A. (1967). On the distribution of the products of random variables. *Journal of the Royal Statistical Society Series B, 29,* 513-523.

MacKinnon, D.P. (2008). *Introduction to Statistical Mediation Analysis.* New York, NY: Taylor & Francis Group.

MacKinnon, D.P., Fritz, M.S., Williams, J., & Lockwood, C.M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods, 39,* 384-389.

MacKinnon, D.P., Goldberg, L., Clarke, G.N., Elliot, D.L., Cheong, J., Lapin, A., … Krull, J.L. (2001). Mediating mechanisms in a program to reduce intentions to use anabolic steroids and improve exercise self-efficacy and dietary behavior. *Prevention Science, 2*, 15-28.

MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7,* 83-104.

MacKinnon, D.P., Lockwood, C. M., and Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39,* 99-128.

Manly, B.F. (1976). Exponential data transformation. *The Statistician, 25,* 37-42.

Oliveira, A., Oliveira, T.A., & Seijas-Macias, A. (2016). Skewness into the product of two normally distributed variables and the risk consequences. *Statistical Journal, 14*, 119-138.

Osborne, J.W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation, 15*(12), 1-9.

R Core Team. (2016). R: A language and environment for statistical computing (Version 3.3.2) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

SAS Institute. (2012). SAS (Version 9.4) [Computer software]. Cary, NC: SAS Inc.

Serlin, R.C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5*, 230-240.

Shrout, P.E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*(4), 422-445.

Springer, M.D., & Thompson, W.E. (1966). The distribution of products of independent random variables. *SIAM Journal of Applied Mathematics, 14*, 511-526.

Tukey, J.W. (1957). The comparative anatomy of transformations. *Annals of Mathematical Statistics, 28*, 602-632.

Zhou, X.H., & Gao, S. (2000). One-sided confidence intervals for means of positively skewed distributions. *The American Statistician, 54*(2), 100-104.

Appendix A. Tables of results

Table A.1

F-values for the method × sample size × a interaction sliced by method for remaining alternatives and comparison groups

Method	*Num df*	*Den df*	*F Value*
Joint significance test	15	156	5.98**
Percentile	15	156	5.09**
z_0	15	156	4.71**
z_{mean}	15	156	4.47**
z_{medi}	15	156	5.09**
z_{g1}	15	156	29.10**
z_{mc}	15	156	4.35**
T_{4p}	15	156	5.09**

** $p < .001$.

Table A.2

F-values for the method × sample size × b interaction sliced by method for the remaining alternatives and the comparison groups

Method	*Num df*	*Den df*	*F Value*
Joint significance test	15	156	3.80**
Percentile	15	156	4.08**
z_0	15	156	3.49**
z_{mean}	15	156	4.02**
z_{medi}	15	156	4.08**
z_{g1}	15	156	31.45**
z_{mc}	15	156	3.70**
T_{4p}	15	156	4.08**

** $p < .001$.

Table A.3

Descriptive statistics for power for each of the remaining methods

Method	*M*	*SD*	*Min*	*Max*
Joint significance test	0.7069	0.3590	0.025	1.000
Percentile	0.7101	0.3552	0.025	1.000
z_0 (bias-corrected)	0.7334	0.3352	0.055	1.000
z_{mean}	0.7326	0.3352	0.056	1.000
z_{medi}	0.7101	0.3552	0.025	1.000
z_{g1}	0.8241	0.2258	0.346	1.000
z_{mc}	0.7311	0.3363	0.054	1.000
T_{1low}	0.9534	0.2109	0.706	1.000
T_{1upp}	0.0389	0.1934	0.000	0.246
T_{2low}	0.9563	0.2045	0.725	1.000
T_{2upp}	0.0423	0.2012	0.000	0.264
T_3	0.9896	0.1017	0.914	1.000
T_{4p}	0.7101	0.3552	0.025	1.000
T_{4z}	0.9889	0.1047	0.910	1.000

Table A.4

Coverage frequency for each of the methods

Method	*Coverage*	*Coverage percentage (%)*	*No-coverage*	*No-coverage percent (%)*
Percentile	60892	95.14	3108	4.86
z_0 (bias-corrected)	60364	94.32	3636	5.68
z_{mean}	60317	94.25	3683	5.75
z_{medi}	60892	95.14	3108	4.86
z_{g1}	52033	81.30	11967	18.70
z_{mc}	60347	94.29	3653	5.71
T_{1low}	34274	53.55	29726	46.45
T_{1upp}	32623	50.97	31377	49.03
T_{2low}	33043	51.63	30957	48.37
T_{2upp}	31214	48.77	32786	51.23
T_3	3888	6.08	60112	93.93
T_{4p}	60892	95.14	3108	4.86
T_{4z}	3971	6.20	60029	93.80

Table A.5

Balance percentage for each of the methods

Method	*Left failure (%)*	*No failure (%)*	*Right failure (%)*
Percentile	2.09	96.30	1.61
z_0 (bias-corrected)	2.67	95.51	1.83
z_{mean}	2.64	95.53	1.83
z_{medi}	2.09	96.30	1.61
z_{g1}	10.73	82.54	6.73
z_{mc}	2.62	95.57	1.81
T_{1low}	46.45	53.55	—
T_{1upp}	—	76.96	23.04
T_{2low}	48.37	51.63	—
T_{2upp}	—	75.81	24.19
T_3	45.69	6.66	47.65
T_{4p}	3.06	95.53	1.42
T_{4z}	48.60	6.96	44.44