Published in final edited form as:
Multivariate Behav Res. 2012 ; 47(1): 61–87. doi:10.1080/00273171.2012.640596.

Explanation of Two Anomalous Results in Statistical Mediation Analysis

Matthew S. Fritz[a], **Aaron B. Taylor**[b], and **David P. MacKinnon**[c]

[a]Virginia Polytechnic Institute and State University

[b]Texas A&M University

[c]Arizona State University

Abstract

Previous studies of different methods of testing mediation models have consistently found two anomalous results. The first result is elevated Type I error rates for the bias-corrected and accelerated bias-corrected bootstrap tests not found in nonresampling tests or in resampling tests that did not include a bias correction. This is of special concern as the bias-corrected bootstrap is often recommended and used due to its higher statistical power compared with other tests. The second result is statistical power reaching an asymptote far below 1.0 and in some conditions even declining slightly as the size of the relationship between *X* and *M*, *a*, increased. Two computer simulations were conducted to examine these findings in greater detail. Results from the first simulation found that the increased Type I error rates for the bias-corrected and accelerated bias-corrected bootstrap are a function of an interaction between the size of the individual paths making up the mediated effect and the 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. Results from the second simulation found that stagnation and decreases in statistical power as a function of the effect size of the a path occurred primarily when the path between *M* and *Y*, *b*, was small. Two empirical mediation examples are provided using data from a steroid prevention and health promotion program aimed at high school football players (Athletes Training and Learning to Avoid Steroids; Goldberg et al., 1996), one to illustrate a possible Type I error for the bias-corrected bootstrap test and a second to illustrate a loss in power related to the size of *a*. Implications of these findings are discussed.

Statistical mediation analysis is used to determine whether the effect of one variable on a second variable is transmitted through a third variable. Mediation analysis is widely used in the social sciences in order to probe how changes in outcome variables occur (MacKinnon, 2008). With the popularity of mediation analysis, numerous statistical tests to determine the presence of mediation have been developed. MacKinnon, Lockwood, Hoffman, West, and Sheets (2002) described 14 different single sample tests of mediation, whereas MacKinnon, Lockwood, and Williams (2004) described an additional 6 resampling tests of mediation.

Correspondence concerning this article should be addressed to Matthew S. Fritz, Department of Psychology, Virginia Tech, 109 Williams Hall, Blacksburg, VA 24061. Matt.Fritz@vt.edu.

Publisher's Disclaimer: The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Several studies have examined the statistical properties of these different tests of mediation for specific situations, including their statistical power, Type I error rates, and coverage of confidence intervals (e.g., Cheung, 2007, 2009; Fritz & MacKinnon, 2007; MacKinnon et al., 2004; MacKinnon, Lockwood et al., 2002; Preacher & Hayes, 2008; Williams & MacKinnon, 2008). Although these studies have been successful in providing advice to researchers as to the accuracy and power of the various tests, they have also produced anomalous findings that are of concern, especially for studies testing mediation with small sample sizes.

In this study, two important anomalous findings are examined. The first is the finding of elevated Type I error rates for two widely recommended resampling tests of mediation: the bias-corrected and accelerated bias-corrected bootstrap tests. The second is the finding that the statistical power for some tests of mediation asymptote at a value below 1.0 and can even decrease as the size of one of the paths increases. The goal of this study is to determine under what conditions these two results occur as well as to estimate their magnitude in order to provide researchers with recommendations on how to avoid inflated Type I error and flattening or declining power when estimating mediational models. Both of these results are then illustrated in empirical examples by estimating mediation models using data from a steroid prevention and health promotion program for high school football players (Athletes Training and Learning to Avoid Steroids; Goldberg et al., 1996), which were selected because of the large sample size and the presence of numerous significant mediation effects (MacKinnon et al., 2001).

STUDY 1

Simulation studies that have investigated the statistical power of the bootstrap methods and the asymmetric confidence interval test using numerical integration with the PRODCLIN program (MacKinnon, Fritz, Williams, & Lockwood, 2007) have found these methods to be more powerful than normal theory tests in many circumstances (Fritz & MacKinnon, 2007; MacKinnon et al., 2004). Specifically, the percentile bootstrap and the numerical integration tests have greater power than normal theory tests in many circumstances while maintaining Type I error near its nominal level. The bias-corrected and accelerated bias-corrected bootstrap tests typically have even greater power than the percentile bootstrap or the numerical integration test. Due to their greater statistical power and the fact that they do not rely on normal theory, many authors have recommended bootstrap methods as the preferred tests of mediation (e.g., MacKinnon et al., 2004; Mallinckrodt, Abraham, Wei, & Russell, 2006; Pituch, Stapleton, & Kang, 2006; Preacher & Hayes, 2008; Shrout & Bolger, 2002).

Given the choice of the different methods available to perform bootstrap tests of mediation, many researchers have naturally chosen to use the bias-corrected bootstrap due to its relatively higher statistical power (e.g., Baker, Pankhurst, & Robinson, 2007; Cerin & Leslie, 2008; Lundgren, Dahl, & Hayes, 2008; Napolitano et al., 2008; Prelow, Weaver, & Swenson, 2006; Tallman, Altmaier, & Garcia, 2007). Furthermore, as cited by Shrout and Bolger (2002), Efron and Tibshirani (1993) suggest that the bias-corrected bootstrap will produce more accurate confidence intervals for small samples. However, in addition to having greater statistical power, the bias-corrected and accelerated bias-corrected bootstrap have both been found to have greatly elevated Type I error rates in some conditions, particularly when the sample size is small (Fritz & MacKinnon, 2007; MacKinnon et al., 2004). This is especially problematic given the possibility that researchers may turn to one of these methods after failing to find significance with another method (e.g., the percentile bootstrap). The conditions under which the bias-corrected and accelerated bias-corrected bootstrap tests of mediation produce elevated Type I error rates have not been explored systematically but need to be examined given these tests' widespread recommendation and frequent usage.

The Single Mediator Model

Mediation occurs when the effect of one variable, the antecedent, on a second variable, the consequent or outcome, is transmitted through a third intervening or mediating variable. Mediation differs from other third variable models such as moderation or suppression in that it explicitly assumes that the variables form a causal chain. That is, changes in the antecedent cause changes in the mediator, which in turn cause changes in the consequent. The single mediator model is the simplest mediation model and is illustrated in Figure 1, where the antecedent is labeled *X*, the mediator is labeled *M*, and the outcome is labeled *Y*. The single mediator model can be represented using three regression equations:

$$Y_i=i_1+c\,X_i+e_i \quad (1)$$

$$M_i=i_2+a\,X_i+e_i \quad (2)$$

$$Y_i=i_3+c'\,X_i+b\,M_i+e_i, \quad (3)$$

where i_1, i_2, and i_3 are intercepts, and e_1 are the residuals. The total effect of *X* on *Y* is represented by *c*, whereas the effect of *X* on *Y* after controlling for the mediator M, called the direct effect, is represented by c'. The effect of *X* on *M*, called the action theory when *X* is directly manipulated, is represented by *a* and the effect of *M* on *Y* after controlling for *X*, called the conceptual theory (Chen, 1990; MacKinnon, Taborga, & Morgan-Lopez, 2002), is represented by *b*. Under the assumption that the mediation model is the correct population model, the mediated effect of *X* on *Y* through *M* is then equal to the indirect path through *M*, *ab*. Sample estimates of *c*, c′, *a*, and *b* are $\hat{c}$, $\hat{c}'$, $\hat{a}$, and $\hat{b}$, respectively.

Tests of Mediation

There are three different categories of tests for examining the single mediator model (MacKinnon, Lockwood et al., 2002). The first category consists of the causal steps approaches first described by Judd and Kenny (1981). In the most widely used version of the causal steps tests, each of the steps in the causal chain is tested in sequence (Baron & Kenny, 1986). First the estimate of the overall effect of *X* on *Y*, $\hat{c}$, is tested for significance, followed by the effect of *X* on *M*, $\hat{a}$, then the effect of *M* on *Y* controlling for *X*, $\hat{b}$. Finally, the comparison $\hat{c}>\hat{c}'$ is made. Although the causal steps test makes sense logically, the causal steps methods do not provide an estimate of the mediated effect and simulation studies have shown the causal steps methods to be underpowered in many situations (Fritz & MacKinnon, 2007; MacKinnon, Lockwood et al., 2002).

The second category of tests is the difference-in-coefficients tests, which assess mediation by dividing $\hat{c}-\hat{c}^{prime}$ by its standard error and comparing the result with a *t*-distribution (MacKinnon, Lockwood et al., 2002). The third category of tests is the product-of-coefficients tests, where the estimate of the indirect effect, $\widehat{ab}$, is divided by its standard error and compared with a normal distribution. MacKinnon, Warsi, and Dwyer (1995) showed that the product-of-coefficients estimator of the mediated effect is equivalent to the difference-in-coefficients estimator: $\hat{c}-\hat{c}'=\widehat{ab}$. Although both of these approaches provide an estimate of the mediated effect and have greater power than the causal steps approaches in many situations, the product-of-coefficients tests are generally preferred in psychology because they generalize more easily to models with multiple mediators. In such models,

$\hat{c} - \hat{c}'$ is an estimate of the total mediated effect, but it cannot be parsed into individual mediated effects passing through the separate mediators. By contrast, if a product-of-coefficients approach is used, $\widehat{ab}$ can easily be calculated and tested for significance separately for each mediator and the estimates of the individual mediated effects can then be summed to form an estimate of the total mediated effect.

The product-of-coefficients tests are not without problems, however. The most concerning issue is their use of an inappropriate sampling distribution. When $\widehat{ab}$ is divided by an estimate of its standard error, for instance the first-order standard error derived by Sobel (1982),

$$\hat{\sigma}^2_{\widehat{ab}} = \hat{b}^2 \hat{\sigma}^2_{\hat{a}} + \hat{a}^2 \hat{\sigma}^2_{\hat{b}}, \quad (4)$$

the resulting statistic is compared with a normal distribution to test for significance, but the distribution of the product of two normally distributed random variables is not itself normally distributed in most situations (Lomnicki, 1967; Springer & Thompson, 1966). In fact, the distribution of the product takes on different shapes for different values of *a, b*, and their respective standard errors. Because of this, the assumption of normality is violated and the results of the *z* tests are inaccurate, as are the symmetric confidence intervals around $\widehat{ab}$ using critical values from the normal distribution.

Two methods have been proposed to deal with the nonnormality of the distribution of the product. The first is a program called PRODCLIN (MacKinnon, et al., 2007), which uses numerical integration to estimate the distribution of the product, given values of $\hat{a}, \hat{b}$, and their standard errors, to find critical values that bound the middle 95% of the distribution (assuming a nominal Type I error rate of $\alpha = .05$). These critical values are then used to form an asymmetric confidence interval around the mediated effect and significance is tested by determining whether or not zero is included within the confidence interval.

The second method is to use bootstrapping, a nonparametric method that does not assume the distribution of *ab* is normal. Bootstrapping creates an empirical sampling distribution of *ab* by resampling many times from the original sample (Bollen & Stine, 1990; Efron & Tibshirani, 1993; MacKinnon et al., 2004; Shrout & Bolger, 2002). Bootstrap samples are the same size as the original sample and they are taken with replacement, meaning that the same case can be selected more than once. For each of the bootstrap samples, $\hat{a}$ and $\hat{b}$ are found and then multiplied to form an estimate of the mediated effect. These $\widehat{ab}$ values for each bootstrap sample collectively make up the bootstrap sampling distribution. The observations corresponding to the $\alpha/2$ and $1 - \alpha/2$ percentiles are the $(1 - \alpha)\times100\%$ confidence limits. Testing the null hypothesis that $ab = 0$ using these confidence limits is the percentile bootstrap test.

The accelerated bias-corrected bootstrap (Efron & Tibshirani, 1993) is a variation on the percentile bootstrap that contains a correction, z_0, for bias in $\widehat{ab}$ and a correction, ν, for changes in the standard deviation of $\widehat{ab}$ as *ab* varies (i.e., acceleration), such that

$$\hat{z}_0 = \Phi^{-1}(\pi), \quad (5)$$

where Φ^{-1} is the inverse function of the normal cumulative distribution function (e.g., $\Phi^{-1}(.975) = 1.96$) and π is the proportion of bootstrap estimates of *ab* that are less than $\widehat{ab}$ from the original data, and

$$\widehat{\nu}=\frac{\sum_{i=1}^{n}\left(\widehat{ab}_{(.)}-\widehat{ab}_{-i}\right)^{3}}{6\left[\sum_{i=1}^{n}\left(\widehat{ab}_{(.)}-\widehat{ab}_{-i}\right)^{2}\right]^{3/2}}, \tag{6}$$

where $\widehat{ab}_{(-i)}$ is the estimate of *ab* with case *i* deleted from the original data set, also called the *i* th jackknife estimate of *ab*, and $\widehat{ab}_{(.)}$ is the mean of all of the jackknife estimates of *ab*. The resulting upper and lower confidence limits are then the bootstrapped estimates of *ab* equal to the percentiles:

$$ABC_{Upper}=\Phi\left(\frac{\widehat{z}_0+z_{(1-\alpha/2)}}{1-\widehat{\nu}\left(\widehat{z}_0+z_{(1-\alpha/2)}\right)}+\widehat{z}_0\right) \tag{7}$$

$$ABC_{Lower}=\Phi\left(\frac{\widehat{z}_0+z_{\alpha/2}}{1-\widehat{\nu}\left(\widehat{z}_0+z_{\alpha/2}\right)}+\widehat{z}_0\right), \tag{8}$$

where Φ is the normal cumulative density function (e.g., Φ(1:96) = .975), $z_{\alpha/2} = \Phi^{-1}(\alpha/2)$, and $z(_{1-\alpha/2}) = \Phi^{-1}(1 - \alpha/2)$. The bias-corrected bootstrap is defined in the same way as the accelerated bias-corrected bootstrap but with $\widehat{\nu}=0$. Note that when both $\widehat{\nu}=0$ and $\widehat{z}_0=0$, the accelerated bias-corrected and bias-corrected bootstraps are equal to the percentile bootstrap.

Methods

In order to investigate what factors affect the Type I error rate of the bias-corrected and accelerated bias-corrected bootstrap tests of mediation, a Monte Carlo simulation was performed using R (R Core Development Team, 2011). Five factors were varied, the first of which was the test of mediation used. In addition to the bias-corrected and accelerated bias-corrected bootstrap tests, the percentile bootstrap and numerical integration tests were used in order to replicate the results from previous simulation studies (MacKinnon et al., 2004) and to examine situations that may increase the Type I error rate for all of these tests.

The second and third factors varied were the sizes of *a* and *b*. Because Type I error rates were being examined, the true mediated effect was 0 in all conditions, requiring that either *a*, *b*, or both be equal to 0. There were seven conditions defined by the combination of the *a* and *b* factors. In three conditions, *b* was set to 0 and *a* was set to 0.14, 0.39, or 0.59; in three other conditions, *a* was set to 0 and *b* was set to 0.14, 0.39, or 0.59; and in one final condition, *a* and *b* were both set to 0. The coefficient sizes of 0.14, 0.39, and 0.59 were chosen to represent small, medium, and large effects (Cohen, 1988). For simplicity, c' was set equal to 0 in all conditions, as Fritz and MacKinnon (2007) found that the size of c' did not have an effect on tests of $\widehat{ab}$. The fourth factor varied was sample size: five sample sizes were selected to cover the range of sample sizes used in the social sciences: 50, 100, 500, 1,000, and 2,500. The final factor varied was the number of bootstrap samples. Although approximately 1,000 bootstrap samples are sufficient for many applications, given the possibility that increasing the number of bootstrap samples would correct for the elevated Type I error rates, the number of bootstrap samples was set at 1,001, 5,001, and 10,001 in different conditions.

The simulation was conducted by first selecting values for *a*, *b*, the sample size, and the number of bootstrap samples. Next, data for *X* was generated from a normal distribution using the RNORM function in R, and data for *M* and *Y* were generated using the selected values for *a* and *b*, and Equations 2 and 3, with residuals added to *M* and *Y* using the

RNORM function. The PRODCLIN program was used to find 95% confidence limits for the numerical integration test (95% confidence limits were used for all four tests because the Type I error level was set at .05). Then, the original sample was bootstrapped the specified number of times and confidence intervals were created using the percentile, bias-corrected, and accelerated bias-corrected bootstrapping methods. Mediation was tested for by examining whether or not each of the confidence intervals contained zero. This process was repeated 1,000 times and the proportion of replications where significant mediation was found by a specific test was that test's Type I error rate. The entire process was repeated for all four tests of mediation within all conditions defined by parameter combinations of *a*, *b*, sample size, and number of bootstrap samples. In total, there were 105 conditions of 1,000 replications each.

The results from the simulation were analyzed using PROC GLM (SAS Institute, 2011). The effects of *a* and *b* were tested using separate ANOVAs because the *a* and *b* factors were not combined factorially in the study due to the requirement that at least one of *a* and *b* must be equal to zero for every condition. Had both effects been estimated in the same ANOVA, the marginal means for levels of *a* and *b* would have been misleading. For example, if *a* and *b* were both entered in the ANOVA, the mean Type I error rate for $a = 0$ would be estimated after collapsing across all four levels of *b*, whereas the mean Type I error rate for the other three values of *a* (i.e., 0.14, 0.39, and 0.59) would only be collapsed across one level of *b* (i.e., $b = 0$). The final two ANOVA models reported here contained all four main effects, all six two-way interactions, and three three-way interactions: type of test by number of bootstrap samples by sample size, type of test by number of bootstrap samples by effect size, and type of test by sample size by effect size. Due to the exploratory nature of the analyses and the potential to test a large number of contrasts to examine the main effects and interactions more thoroughly, elevated Type I error rates were a concern. Hence, an effect size threshold of $\eta^2 \geq .01$ was chosen for interpreting effects and null hypotheses corresponding to all η^2 values reaching this threshold were rejected at $p < .001$.

Results

The results showed that for $b = 0$ there were significant main effects of the effect size of *a*, F(3, 24) = 164.34, $\eta^2 = .677$, and test, F(3, 72) = 420.24, $\eta^2 = .050$; note that all reported effect sizes are semi-partial η^2 values. When $a = 0$, there were also main effects of the effect size of *b*, F(3, 24) = 443.35, $\eta^2 = .675$, and test, F(3, 72) = 508.22, $\eta^2 = .055$. No main effects of number of number of bootstrap samples or sample size were found for either model. The significant main effects for test and effect size were further investigated using contrasts. For the main effect of test, the bias-corrected and accelerated bias-corrected bootstrap tests together had higher Type I error rates than the numerical integration and percentile bootstrap tests together when $b = 0$ and when $a = 0$, F(1, 24) = 710.89, $\eta^2 = .049$ and F(1,24) = 1678.74, $\eta^2 = .054$, respectively. However, the Type I error rate did not significantly differ between the bias-corrected and the accelerated bias-corrected bootstrap or between the percentile bootstrap and the numerical integration test for either model. The main effect of effect size was tested for linear and quadratic trends, which were found to be significant, F(1, 24) = 408.04, $\eta^2 = .560$ and F(1, 24) = 84.82, $\eta^2 = .117$, respectively, for *a* and F(1, 24) = 1127.17, $\eta^2 = .572$, F(1, 24) = 183.87, $\eta^2 = .093$, respectively, for *b*. As shown in Figure 2, for both *a* and *b*, moving from 0 to 0.14 and then to 0.39 greatly increased the Type I error, but moving from 0.39 to 0.59 only marginally increased the Type I error.

The results showed that no interactions that contained number of bootstrap samples were significant. The interaction between effect size and sample size was significant F(12, 24) = 10.74, $\eta^2 = .177$, for *a* and F(12, 24) = 30.17, $\eta^2 = .184$, for *b*, and was quite large in comparison to all of the other effects in the model except the main effect of effect size, indicating that the effect of effect size on the Type I error rate differed depending on the sample size. The interactions between test and effect size, and test and sample size were either significant or approaching significance for both models, so a three-way interaction between test, sample size, and effect size was added to the model. The three-way interaction was significant for both models, F(36, 72) = 13.81, $\eta^2 = .020$, for *a* and F(36, 72) = 13.19, $\eta^2 = .017$, for *b*. The three-way interaction for a is illustrated in Figure 3 which shows the Type I error for the four tests averaged across the number of bootstrap samples when $b = 0$. As indicated by the interaction, the effect of sample size on the Type I error differed depending on the effect size of the non-zero path, and this differential effect differed depending on the test used. As shown by the figure, the conditions where *a* and *b* were both equal to zero had very low Type I error for all tests. For the bias-corrected and accelerated bias-corrected bootstrap, when *a* was 0.39 or 0.59. and the sample size was smaller than 500, the Type I error rate was elevated, approaching or even exceeding .09, but as the sample size approached 2500, the Type I error rate returned to .05. This effect was much less pronounced for the percentile bootstrap and numerical integration tests, with the Type I error rate never exceeding .062. When $a = 0.14$, elevated Type I errors occurred for sample sizes of 500 and 1000 for the bias-corrected and accelerated bias-corrected bootstraps, but not for the other two tests. A second figure is not provided for the effect of b as it was very similar to Figure 3.

In addition to the Type I error rate, the number of times that the confidence interval failed to the left or right for each test (i.e., how often the upper confidence limit was less than the true value of zero and how often the lower confidence limit was greater than the true value) was also recorded to determine if any of the tests failed more in a specific direction. No significant main effects were found in the balance of the failure rates for any of the four tests. As a result, no moderating effects of sample size or effect size were tested.

Discussion

As found in simulations conducted by MacKinnon et al. (2004) and Cheung (2007), the bias-corrected and accelerated bias-corrected bootstrap tests of mediation were found to have elevated Type I error rates for certain conditions, whereas the percentile bootstrap and numerical integration tests were found to have Type I error rates closer to .05 for those conditions. The number of bootstrap samples did not have an effect on the Type I error rate for any of the tests in this study, indicating that simply increasing the number of bootstrap samples will not correct the elevated Type I error rates.

The size of the nonzero path did have a significant effect on the Type I error rate for all of the tests. This is not surprising because we are calculating the Type I error rate of the product $\widehat{ab}$. In the situation where $b = 0$ and $a > 0$, it is likely that even though $b = 0$, $\widehat{b}$ will not be exactly equal to zero, meaning that $\widehat{ab}$ will not be exactly equal to zero either. As *a* increases, the probability that $\widehat{ab}$ will be significant increases. Looking at Figure 2, the curvilinear relationship shows that this effect is quite strong as *a* ranges from 0 to 0.39 but weakens for larger effects, indicating that there are only a few instances where $\widehat{ab}$ would be significant when $a = 0.59$ that would not also be significant if $a = 0.39$ for the same value of $\widehat{b}$ This same pattern holds true when $a = 0$ and $b > 0$.

The most interesting result from the simulation is the significant interaction between test, sample size, and effect size. As illustrated in Figure 3, when $a > 0.14$, the bias-corrected and accelerated bias-corrected bootstrap had elevated Type I error rates for sample sizes less than 500, but these approached .05 as the sample size increased beyond 500. The percentile bootstrap and numerical integration tests also had elevated Type I error rates for sample sizes less than 500, but the effect was much less pronounced and the Type I error never exceeded .062 for either test. When $a = 0.14$, the Type I error rates for the bias-corrected and accelerated bias-corrected bootstrap tests were conservative for sample sizes less than 500, elevated for samples sizes between 500 and 1,000, and approached .05 as the sample size approached 2,500. In contrast, the percentile bootstrap and numerical integration tests had Type I error rates that were conservative for sample sizes of 500 or less. When b > 0, the same pattern was produced.

The results suggest that when a study uses a sample of less than 500 from a population where either a or b is small, the other effect is zero, Type I errors are rare, regardless of the test used. If either a or b is small, the other effect is zero, and the sample size is in the range of 500 to 1,000, then using the bias-corrected or the accelerated bias-corrected bootstrap results in an elevated number of Type I errors compared with the percentile bootstrap and the numerical integration tests. This is due to these tests becoming adequately powered to find small mediated effects as the sample size approaches 500 as found by Fritz and MacKinnon (2007). If either a or b is medium or large, the other effect is zero, and the sample size is less than 500, using the bias-corrected or accelerated bias-corrected bootstrap results in elevated Type I errors, in some cases almost twice as high as the nominal level. The percentile bootstrap and numerical integration tests also have elevated Type I errors for these conditions but less so than the other two tests.

In general, once the sample size reaches 2,500, Type I error rates for all of the tests approach .05, regardless of the effect size of the nonzero path. This would indicate that the bias-corrected and accelerated bias-corrected bootstrap tests are overcorrecting for small sample sizes, which is likely due to problems with estimating the bias correction term, z_0, for small samples. As Efron and Tibshirani (1993, p. 327) noted, not only is z_0 an estimate of the median bias, not the mean bias, but also it is actually easier to find a good estimate of ν than z_0. Because of this, it may be that $\hat{z}_0$ is inflated for small samples and a better estimate of bias is needed. However, a different estimator of bias may reduce the gain in statistical power of the bias-corrected bootstrap tests over the uncorrected percentile bootstrap test.

Empirical Example

In order to illustrate how the increased Type I error rates found for the bias-corrected bootstrap in Study 1 could affect the results from a mediation analysis, we present an empirical example. The data for the empirical example comes from the Athletes Training and Learning to Avoid Steroids (ATLAS; Goldberg et al., 1996) study, which is a program designed to reduce high school football players' use of anabolic steroids by presenting healthy nutrition behaviors and strength training as direct alternatives to steroid use. MacKinnon et al. (2001) investigated 12 possible mediators of the ATLAS program on the three outcome variables of interest: intentions to use anabolic steroids, nutrition behaviors, and strength training self-efficacy. In the current example, the model tested whether a student's perceived susceptibility to the adverse effects of steroid use, measured immediately after completion of the ATLAS program, was a mediator of the relationship between participation in the ATLAS program and the student's intentions to use anabolic steroids measured 9 months after completion of the ATLAS program.

A random sample of 700 freshman and junior students was selected from Cohort 1 of the ATLAS study for the analysis. Ninety-nine students were removed due to missing values on both the mediator and the outcome, leaving a final sample of 601 students. When the data were analyzed using Mplus (Muthen & Muthen, 2011), the bias-corrected bootstrap 95% confidence interval (CI) was [−0.085, −0.001], indicating that perceived susceptibility was a significant mediator of the ATLAS program and intentions to use steroids. However, the percentile bootstrap 95% CI was [−0.079, 0.002] and the normal theory 95% CI was

[−0.067, 0.010], indicating that susceptibility was not a significant mediator. In addition, the PRODCLIN 95% CI was [−0.07461, 0.00026], also nonsignificant. These results indicate that using the bias-corrected bootstrap to test mediation in this example could result in a Type I error. It is also possible that the bias-corrected bootstrap is correct whereas the other tests are underpowered, meaning that using tests other than the bias-corrected bootstrap would result in a Type II error. However, Fritz and MacKinnon (2007) found that the power differences are not that large and for a sample size of 601, all of these tests should be adequately powered to find even small effects. Of course, given that these are real data, the true presence or absence of mediation remains unknown. In this circumstance, when multiple tests of the same effect return conflicting results, unless there is a specific reason one test should be chosen over another, it would seem the safest course of action would be to take the conservative approach and report a nonsignificant finding.

STUDY 2

A second anomalous result that arises for tests of mediation is that the power to detect the mediated effect can asymptote at a value far below 1.0 or even decline as effect size increases (Taylor, MacKinnon, & Fritz, 2007; similar results were found but unreported by Fritz & MacKinnon, 2007). Specifically, this stagnation or reduction in power occurs when *a* increases. The reason is most clearly seen when considering testing mediation using a causal steps approach. In the model in which the outcome variable is regressed on the antecedent and the mediator (i.e., Equation 3), *a* no longer represents an effect of interest but instead is the correlation between the antecedent and the mediator. As *a* increases, and therefore the correlation between *X* and *M* increases, $\widehat{b}$, which is the effect of interest in this model, has its standard error increase (Hoyle & Kenny, 1999; Shrout & Bolger, 2002). It is therefore possible to have a net loss of statistical power to detect the mediated effect if the increase in the standard error of $\widehat{b}$ reduces the power to reject H_0 for *b* more than the increase in *a* increases power to reject H_0 for *a*. This is the standard error formula for $\widehat{b}$ in the case where the antecedent, the mediator, and the outcome are all standardized (modified from Cohen, Cohen, West, & Aiken, 2003, Equation 3.6.2).

$$\widehat{s_{\widehat{b}}}=\sqrt{\frac{1-\left(\widehat{b}^2+\widehat{c'}^2+2\widehat{a}\widehat{b}\widehat{c'}\right)}{(n-3)\left(1-\widehat{a}^2\right)}}. \quad (9)$$

Note that $\widehat{a}$ appears in the denominator, so an increase in $\widehat{a}$ will reduce the denominator and therefore increase the standard error, reducing power to reject H_0 for *b*. There is no corresponding effect of the size of $\widehat{b}$ on the standard error of $\widehat{a}$ because $\widehat{a}$ is estimated using a model including only the antecedent and the mediator (i.e., Equation 2).

The problem of power stagnation and decline is unexpected because other statistical methods typically have power that increases monotonically as effect size increases and asymptotes at 1.0 rather than at a lower value. This study investigates under what circumstances this power stagnation and decline occurs. A second purpose of this study is to assess whether the stagnation and decline in statistical power occur for methods not using the causal steps approach. As bootstrap approaches do not require calculation of the standard error of $\widehat{b}$, whereas the numerical integration test does, it is an open question whether they are susceptible to the power stagnation and decline problem.

Method

A Monte Carlo simulation was performed to discover under what conditions increases in the size of *a* are associated with power stagnation and decline and for which tests of mediation this occurs. Data were simulated and analyzed using SAS (SAS Institute, 2011). All three variables, the antecedent, the mediator, and the outcome, were simulated to be normally distributed and standardized in the population. Four between-conditions factors were manipulated in the study: sizes of *a*, *b*, and c', and sample size, *n*. The true values of *a* and *b* were each varied from .02 to .60 in increments of .02. (True values of 0 were not studied because rejections of the null hypothesis when $ab = 0$ would constitute a Type I error rather than a correct rejection.) The true value of c' was set to either 0 or .50. These values were chosen to bracket the range of likely effects; previous research (Fritz & MacKinnon, 2007; MacKinnon, Lockwood et al., 2002) has shown that tests of mediation, other than the causal steps tests, are largely unaffected by the size of c'. The upper limits of the sizes of these three coefficients were constrained by the standardization of the variables. Because all variables were standardized, the values of *a*, *b*, and c' are not only regression coefficients but also correlation coefficients. If they had taken on any larger sizes, this would have resulted in impossible patterns of correlations (i.e., $R_Y^2 > 1$). The final between-conditions factor was sample size, which ranged from 20 to 500 in increments of 20. Six different tests of mediation constituted a within-condition factor: Baron and Kenny's (1986) test, a test using the first-order standard error (Sobel, 1982), the joint significance test (James & Brett, 1984; MacKinnon et al., 2002), the numerical integration test (MacKinnon, Lockwood et al., 2007), the percentile bootstrap, and the bias-corrected bootstrap (Efron & Tibshirani, 1993). One thousand replications were used for each condition.

The effect of interest was to be found not simply in statistical power, or the proportion of replications in a condition in which the null hypothesis was rejected, but in a power curve, or the *changes* in power across conditions representing different values of *a*. Results of the study for each test were therefore aggregated first into power values in each condition and then combined across values of *a* into power curves showing power at levels of *a* in each condition defined by the other design factors (*b*, c', sample size, and test). For each power curve, stagnation was measured as a single variable that also encompassed decline. It was found by comparing the observed power curve with an ideal power curve for that condition. Ideal power curves were constructed using logistic regression because the ogive shape of predicted probabilities it produces matches well the often-seen ogive shape of power curves. The ideal logistic curve was found by first locating the point of maximum slope in the observed power curve for any three consecutive points and then fitting a logistic regression model to the power values to the left of that point (i.e., for smaller values of *a*). In a logistic model, the two halves of the curve divided at the point of maximum slope are mirror images of each other. This means that the ideal curve's standard of comparison for the upper half of the power curve was a mirror image of the observed lower half of the power curve. Figure 4 shows the observed and ideal power curves for the numerical integration test for $b = .20$, $c' = .50$, and $n = 160$. The maximum slope in the actual power curve occurs at $a = .16$. A logistic regression model was fit to the actual power curve for values of *a* up to .16 and was then extended to make the ideal curve. Stagnation was taken as the sum of the differences between the ideal and the actual power curve at all points (values of *a*) to the right of the point of maximum slope. (Negative differences, where the actual power curve was higher than the ideal power curve, were rare and small and were therefore ignored.) Stagnation values were scaled to be proportions of the area of the entire figure (with *a* ranging from 0 to .60 and power ranging from 0 to 1). In Figure 4, stagnation = .104. This can be interpreted as the average amount that the actual power curve falls short of the ideal power curve across the entire range of *a*.

Results

Power stagnation values were analyzed using a test × sample size × $b \times c'$ ANCOVA. Although *a* was manipulated in the study, it was not included in the models because calculating power stagnation required first aggregating power levels across levels of *a* to get a power curve. Additionally, *b* was treated as a continuous predictor in the ANCOVA because its 30 levels were too many to allow the model to be estimated with all possible interactions when it was entered as a categorical predictor. This approach traded the multiple degree of freedom test of any possible form of association between *b* and power stagnation for a single degree of freedom test for a linear association. Because the sample size was so large—9,000 conditions even after aggregating replications within conditions to find estimated power values and aggregating conditions across values of a to make power curves —effect sizes were used in place of significance tests to determine which effects were large enough to be interpretable. As in Study 1, a threshold of $\eta^2 \geq .01$ was chosen for interpreting effects and null hypotheses corresponding to all η^2 values reaching this threshold were rejected at $p < .001$.

In the ANCOVA, the largest effect by far was for *b* ($\eta^2 = .463$). Smaller values of *b* were associated with more power stagnation (see Figure 5). The sample size effect also reached the threshold ($\eta^2 = .066$): power stagnation was lowest in the smallest samples, increased with increasing sample size to about $n = 300$, and then declined slightly. The interaction of these two factors also reached the threshold ($\eta^2 = .051$). As shown in Figure 5, for the smallest samples (≤ 100), power stagnation was fairly low and constant, declining only gradually as *b* increased. As sample size increased, power stagnation increased for smaller values of *b*, and the drop-off in stagnation with increasing *b* became more dramatic. The main effect of test also reached the threshold ($\eta^2 = .014$). Power stagnation was greatest for the joint significance test, the numerical integration test, and the percentile bootstrap (see Figure 6). The test × *b* interaction also reached the threshold ($\eta^2 = .012$). As shown in Figure 6, power stagnation was greatest for the numerical integration test (as well as the joint significance and percentile bootstrap tests, which are not shown in the figure because their results were very similar to those for the numerical integration test) followed by the bias-corrected bootstrap and the first-order standard error. The tests generally had a peak of power stagnation at a value of *b* less than .10 and declining power stagnation with increasing *b*. The Baron and Kenny (1986) test performed quite differently depending on the value of c $'$. When $c' = .5$, the Baron and Kenny test had the highest peak stagnation and the steepest decline of any test. When $c' = 0$, on the other hand, it had very little power stagnation, and its stagnation declined slowly with increasing values of *b*. This difference in performance by value of c' resulted in two interactions involving c' reaching the threshold: test × c' ($\eta^2 = .024$) and test × b × c' ($\eta^2 = .019$). When the ANCOVA was reestimated with the Baron and Kenny test excluded, no effects involving c' reached the interpretation threshold.

Visual inspection of the power curves being analyzed suggested that overall power might mediate the effects of the design factors on power stagnation. In conditions where power was generally quite low, power curves did not increase to anywhere near 1.0 even with the largest values of *a*, and power stagnation was near zero. These power curves simply did not reach high enough values for stagnation to occur. At the other extreme, in conditions where power was high, power curves increased quickly as *a* increased, to asymptote at 1.0, closely matching the corresponding ideal power curves and declining little if at all. Power stagnation appeared most prevalent in conditions of moderate power, where power increased with *a*, but not too steeply.

A joint significance approach was used to test the hypothesis that overall power mediated the relationship between the design factors and power stagnation. As in the previous ANCOVAs, η^2 was used in place of significance tests for interpretation. The first step was an ANCOVA predicting overall power using test, sample size, *b*, and c'; *b* was again entered as a continuous predictor to make the design matrix manageable. Overall power was defined as the proportion of the total area in the figure falling under the power curve, which means it may also be interpreted as average power collapsing across values of *a*. This ANCOVA accounted for the majority of the variance in average power ($\eta^2 = .797$), confirming the association between the design factors and average power. The second step was to conduct another ANCOVA predicting power stagnation from the study design factors and average power. Average power was entered only as a main effect because the question of interest was whether it would have an effect on power stagnation over and above the effects of the study design factors rather than whether it would interact with them. Because the effect of average power was expected to be nonmonotonic—low stagnation at both low and high average power and greater stagnation at moderate average power—it was entered as a categorical predictor. This allowed the modeling of different shapes of association with power stagnation rather than looking for only a linear effect. Average power was entered using 10 categories in increments of .10 (power ≤ .10, .10 < power ≤ .20, etc.). In the ANCOVA for power stagnation, average power accounted for the most variance of any effect ($\eta^2 = .339$; overall $\eta^2 = .930$). As expected, power stagnation increased with increasing average power up to about .25 and then decreased to near zero (see Figure 7). The association between average power and power stagnation suggests that average power did in fact mediate the relation between the study factors and power stagnation.

Discussion

The purpose of this study was to discover under what circumstances increasing a is associated with stagnating and declining power to detect the mediated effect. The size of b was found to be by far the largest single predictor. Power stagnation reached its maximum for b values no larger than .10 and diminished with increasing b. This effect of b was qualified by interactions with sample size and test, indicating that peak stagnation varied with other study factors, but the overall pattern of diminishing power stagnation with increasing b remained. Further analyses showed that this association was mediated by overall power. In combinations of study design factors that led to low average power (lower power tests, smaller sample sizes, etc.), power stagnation was low. In conditions in which average power was moderate, power stagnation increased, but as average power increased, power stagnation diminished to near zero.

The finding that increasing a does not always increase power (and occasionally even reduces it) has important implications for detecting mediated effects. In the common situation where X is an intervention, Y is an outcome, and M is a mediator targeted by the intervention, these results suggest that it may be more important to focus on the association between the mediator and the outcome (conceptual theory) than on the association between the intervention and the mediator (action theory). If the conceptual theory is strong enough, in other words if b is large enough, then power stagnation is near zero. If the conceptual theory is weaker (if b is smaller), then strengthening the action theory (attempting to increase a) by refocusing the intervention to more precisely target the mediator may not improve power or may even be counterproductive. In the process of choosing potential mediators of an intervention, then, it seems wisest to focus first on the expected causal link between the mediator and the outcome and only later on the ease with which they can be modified by the intervention. This would be particularly important in circumstances where expected power to detect the mediated effect is not high (e.g., limited sample size).

Empirical Example

As for Study 1, we now present an empirical example for Study 2 using data from the ATLAS program (Goldberg et al., 1996) to illustrate the potential impact that the size of $\widehat{a}$ can have on a mediation analysis when the size of $\widehat{b}$ remains constant. The mechanism of interest for this example is whether a student's knowledge of the adverse effects of anabolic steroid use mediates the relationship between participation in the ATLAS program and intentions to use anabolic steroids. Two models were estimated for this example. In Model

1, knowledge was measured immediately after completion of the ATLAS program and intentions were measured 9 months after completion of the program. In Model 2, knowledge was measured 9 months after completion of the ATLAS program and intentions were measured 12 months after completion of the program. The reason these two models were selected is that the effect of the ATLAS program on knowledge was expected to diminish with time (i.e., the *a* path should be smaller in Model 2) whereas the relationship between knowledge and intentions was not expected to diminish over time (i.e., the *b* paths should be approximately the same in Model 1 and Model 2).

Equations 2 and 3 were estimated for Models 1 and 2 using PROC REG (SAS Institute, 2011) with a random sample of 279 students who had complete data for the five variables of interest. As expected, $\widehat{a}$ for Model 1 was larger than for Model 2, as shown in Table 1, whereas the $\widehat{b}$ paths for both models were approximately the same size. What is important to note, however, is that even though these models were estimated using the same number of cases, the size of the $\widehat{b}$ paths were equivalent, and the squared multiple correlation for Equation 3 for Model 1 was larger than the squared multiple correlation for Equation 3 for Model 2, the standard error for the $\widehat{b}$ path in Model 2 was smaller than for Model 1. This means that the power to detect the *b* effect was greater in Model 2, where $\widehat{a}$ was smaller. One could argue that this effect is negligible because $\widehat{b}$ was significant for both models. However, as the difference in the magnitude of the *a* paths between the two models increases, the stagnation of power will continue to increase as well, increasing the probability of making a Type II error.

GENERAL DISCUSSION

The results from the two studies presented here provide a wider insight into the testing of mediational relationships by systematically examining two anomalous findings from previous studies of tests of mediation that have hereto been unexplored. The issue of elevated Type I error rates is informative because although it is already well known that tests of mediation which take into account the nonnormality of the indirect effect are more accurate and more powerful than methods that do not, these results suggest, at least in part, why the bias-corrected and the accelerated bias-corrected bootstrap tests are more powerful than the percentile bootstrap and numerical integration tests. The results of this study indicate that these tests' increased power may simply be a result of their generally higher rejection rate, which also leads to their increased Type I error. If the increases in statistical power are due solely to an elevation in the Type I error, then introducing a Type I error correction to the bias-corrected bootstrap methods would result in the loss of the power gain over the percentile bootstrap. There are other factors that come into play regarding the increased power and elevated Type I error of the bias-corrected methods, however, such as sample size and the size of the nonzero path, so a simple Type I error correction does not appear to be a cure-all solution.

The issue of stagnating statistical power for the *b* effect as the size of the *a* effect increases is also informative as it differs from the more common situation where statistical power always increases as the effect size increases. As described earlier, the effect on power of a large *a* effect was generally seen only when the *b* effect is 0.10 or smaller or overall power of the test was low. One possible explanation for this finding is that for tests of mediation that examine the indirect effect rather than the individual paths, when $b \geq 0.10$ the power to find the *a* effect is driving the power of the test, but when $b \leq 0.10$, the *b* effect is limiting the power of the test. Regardless, the redundancy of *X* and *M* needs to be taken into account when determining the statistical power for the *b* path.

Together, the results from the two studies described here present a more complete picture of the relationship between statistical power, Type I error rate, effect size, and sample size for tests of mediation than what has previously been presented in studies of the performance of statistical tests of mediation. Perhaps the most important message is that the idea that as one of these quantities is increased while the other two are held constant, the fourth quantity must increase is not always as straightforward as it appears because it is not always easy to hold the other two quantities constant. The increases in power for the bias-corrected bootstrap methods come with an increase in Type I error and the increased power that should accompany an increased effect size is tempered by the increased redundancy between *X* and *M*, which in turn increases the standard error of the *b* path.

RECOMMENDATIONS

The findings in this study have two major implications for researchers testing for mediation. The first implication is that although the bias-corrected and accelerated bias-corrected bootstrap have repeatedly been found to have the highest statistical power of the common tests of mediation, and thus are often described as the “best” tests of mediation, we caution researchers about equating “most powerful” with “best” or “most accurate.” Indeed, as shown in Figure 3, the percentile bootstrap and numerical integration tests are more accurate in terms of Type I error than either of the bias-corrected bootstrap tests when sample sizes are small. The decision as to which test to use then becomes one of which is more desirous for a specific research question: minimizing the probability of making a Type I error or minimizing the probability of making a Type II error (i.e., maximizing power). In general, we recommend that researchers choose a single statistical test a priori to test a particular indirect effect, based on whether avoiding Type I or Type II error is of greater concern. We do not recommend that researchers use multiple statistical tests to test the same indirect effect because if the tests disagree, as in the empirical example for Study 1, it is impossible with real data to tell which one is correct, and the obvious temptation will be to simply report whichever test supports the researcher's hypothesis. Instead, in addition to the chosen test of the indirect effect, we recommend examining the significance of $\widehat{a}$ and $\widehat{b}$ individually. The significance tests of these coefficients are unaffected by the nonnormality of the indirect effect. If these effects are both significant, it lends credence to the significance of the test of the indirect effect and vice versa.

The second implication is given that increasing the association between the antecedent and the mediator may fail to improve power to detect the mediated effect for small sample sizes, we recommend that, if forced to choose, researchers with smaller samples should focus on selecting mediators that have a stronger possible relationship with the outcome than with the antecedent. This finding is especially relevant in the case of an intervention that has been specifically designed to change the mediator because most researchers design interventions to have maximal effects on the mediator in order to have maximal effects on the outcome variables. If the intervention is successful and results in a large *a* path, whereas the *b* path, which is not directly manipulated, is a medium or small effect, then the success of the intervention in changing the mediator may have actually decreased the researcher's ability to find this effect by decreasing the statistical power. If the mediator is measured with error, the power to detect the *b* effect is even further reduced, exacerbating the issue (Hoyle & Kenny, 1999). This is not to suggest that researchers should not consider mediators which have large *a* paths. Instead, if mediators with large *a* paths are likely to occur, be certain that the study is powered adequately to counteract the power stagnation that may occur.

In summary, the results of the studies reported in this article suggest that when *a* is a large effect, the statistical power for tests of mediation may be less than is expected, so when *a* is expected to be large, researchers may need to overpower their studies (in terms of increasing sample size, for example) in order to counteract this effect. The results also suggest that for sample sizes less than 2,500, the percentile bootstrap and numerical integration tests have more accurate Type I error rates compared with the bias-corrected and accelerated bias-corrected bootstrap tests. Based upon the results from Fritz and MacKinnon (2007), the increased probability of making a Type I error may not be worth the increase in statistical power, especially as the difference in power between the different tests of mediation decreases as the effect sizes of *a* and *b* increase.

Acknowledgments

Portions of this work were presented at the annual meeting of the Society for Prevention Research. This research was supported in part by a grant from the National Institute on Drug Abuse (DA 009757).

REFERENCES

Baker SR, Pankhurst CL, Robinson PG. Testing relationships between clinical and non-clinical variables in xerostomia: A structural equation model of oral health-related quality of life. Quality of Life Research. 2007; 16:297–308. doi: 10.1007/s11136-006-9108-x. [PubMed: 17033902]

Baron RM, Kenny DA. The moderator-mediation variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology. 1986; 51:1173–1182. doi: 10.1037/0022-3514.51.6.1173. [PubMed: 3806354]

Bollen KA, Stine RA. Direct and indirect effects: Classical and bootstrap estimates of variability. Sociological Methodology. 1990; 20:115–140. doi: 10.2307/271084.

Cerin E, Leslie E. How socio-economic status contributes to participation in leisure-time physical activity. Social Science & Medicine. 2008; 66:2596–2609. doi: 10.1016/j.socscimed.2008.02.012. [PubMed: 18359137]

Chen, H-T. Theory-driven evaluations. Sage; Newbury Park, CA: 1990.

Cheung MW-L. Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. Structural Equation Modeling. 2007; 14:227–246. doi: 10.1080/10705510701293627.

Cheung MW-L. Comparison of methods for constructing confidence intervals of standardized indirect effects. Behavior Research Methods. 2009; 41:425–438. doi: 10.3758/BRM.41.2.425. [PubMed: 19363183]

Cohen, J. Statistical power analyses for the behavioral sciences. 2nd ed.. Erlbaum; Mahwah, NJ: 1988.

Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd ed.. Erlbaum; Mahwah, NJ: 2003.

Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Chapman & Hall/CRC; Boca Raton, FL: 1993.

Fritz MS, MacKinnon DP. Required sample size to detect the mediated effect. Psychological Science. 2007; 18:233–239. doi: 10.1111/j.1467-9280.2007.01882.x. [PubMed: 17444920]

Goldberg L, Elliot D, Clarke GN, MacKinnon DP, Moe E, Zoref L, Lapin A. Effects of a multidimensional anabolic steroid prevention intervention: The Adolescents Training and Learning to Avoid Steroids (ATLAS) program. Journal of the American Medical Association. 1996; 276:1555–1562. doi: 10.1001/jama.276.19.1555. [PubMed: 8918852]

Hoyle, RH.; Kenny, DA. Sample size, reliability, and tests of statistical mediation. In: Hoyle, R., editor. Statistical strategies for small sample research. Sage; Thousand Oaks, CA: 1999. p. 195-222.

James LR, Brett JM. Mediators, moderators, and tests for mediation. Journal of Applied Psychology. 1984; 69:307–321. doi: 10.1037/0021-9010.69.2.307.

Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. Evaluation Review. 1981; 5:602–619. doi: 10.1177/0193841X8100500502.

Lomnicki ZA. On the distribution of the products of random variables. Journal of the Royal Statistical Society Series B. 1967; 29:513–523.

Lundgren T, Dahl J, Hayes SC. Evaluation of mediators of change in the treatment of epilepsy with acceptance and commitment therapy. Journal of Behavioral Medicine. 2008; 31:225–235. doi: 10.1007/s10865-008-9151-x. [PubMed: 18320301]

MacKinnon, DP. Introduction to statistical mediation analysis. Erlbaum; New York, NY: 2008.

MacKinnon DP, Fritz MS, Williams J, Lockwood CM. Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. Behavior Research Methods. 2007; 39:384–389. doi: 10.3758/BF03193007. [PubMed: 17958149]

MacKinnon DP, Goldberg L, Clarke GN, Elliot DL, Cheong J, Lapin A, Krull JL. Mediating mechanisms in a program to reduce intentions to use anabolic steroids and improve exercise self-efficacy and dietary behavior. Prevention Science. 2001; 2:15–28. doi: 10.1023/A:1010082828000. [PubMed: 11519372]

MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychological Methods. 2002; 7:83–104. doi: 10.1037/1082-989X.7.1.83. [PubMed: 11928892]

MacKinnon DP, Lockwood CM, Williams J. Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivariate Behavioral Research. 2004; 39:99–128. doi: 10.1207/s15327906mbr3901_4. [PubMed: 20157642]

MacKinnon DP, Taborga MP, Morgan-Lopez AA. Mediation designs for tobacco prevention research. Drug and Alcohol Dependence. 2002; 68:S69–S83. doi: 10.1016/S0376-8716(02) 00216-8. [PubMed: 12324176]

MacKinnon DP, Warsi G, Dwyer JH. A simulation study of mediated effect measures. Multivariate Behavioral Research. 1995; 30:41–62. doi: 10.1207/s15327906mbr3001_3. [PubMed: 20157641]

Mallinckrodt B, Abraham WT, Wei M, Russell DW. Advances in testing the statistical significance of mediation effects. Journal of Counseling Psychology. 2006; 53:372–378. doi: 10.1037/0022-0167.53.3.372.

Muthen, LK.; Muthen, BO. Mplus user's guide. 6th ed.. Author; Los Angeles, CA: 2011.

Napolitano MA, Papandonatos GD, Lewis BA, Whitely JA, Williams DM, King AC, Marcus BH. Mediators of physical activity behavior change: A multivariate approach. Health Psychology. 2008; 27:409–418. doi: 10.1037/0278-6133.27.4.409. [PubMed: 18642998]

Pituch KA, Stapleton LM, Kang JY. A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. Multivariate Behavioral Research. 2006; 41:367–400. doi: 10.1207/s15327906mbr4103_5.

Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods. 2008; 40:879–891. doi: 10.3758/BRM.40.3.879. [PubMed: 18697684]

Prelow HM, Weaver SR, Swenson RR. Competence, self-esteem, and coping efficacy as mediators of ecological risk and depressive symptoms in urban African American and European American youth. Journal of Youth and Adolescence. 2006; 35:507–517. doi: 10.1007/s10964-006-9068-z.

R Core Development Team. R (Version 2.13.0) [Computer software]. R Foundation for Statistical Computing; Vienna, Austria: 2008. Retrieved from http://www.R-project.org

SAS Institute. SAS (Version 9.2) [Computer software]. SAS Inc; Cary, NC: 2011.

Shrout PE, Bolger N. Mediation in experimental and nonexperimental studies: New procedures and recommendations. Psychological Methods. 2002; 7:422–445. doi: 10.1037/1082-989X.7.4.422. [PubMed: 12530702]

Sobel ME. Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology. 1982; 13:290–312. doi: 10.2307/270723.

Springer MD, Thompson WE. The distribution of products of independent random variables. SIAM Journal of Applied Mathematics. 1966; 14:511–526. doi: 10.1137/0114046.

Tallman BA, Altmaier E, Garcia C. Finding benefit from cancer. Journal of Counseling Psychology. 2007; 54:481–487. doi: 10.1037/0022-0167.54.4.481.

Taylor, AB.; MacKinnon, DP.; Fritz, MS. Power curves for mediation. Poster presented at the annual meeting of the Society for Prevention Research; Washington, DC. Jun. 2007

Williams J, MacKinnon DP. Resampling and distribution of the product methods for testing indirect effects in complex models. Structural Equation Modeling. 2008; 15:23–51. doi: 10.1080/10705510701758166. [PubMed: 20179778]

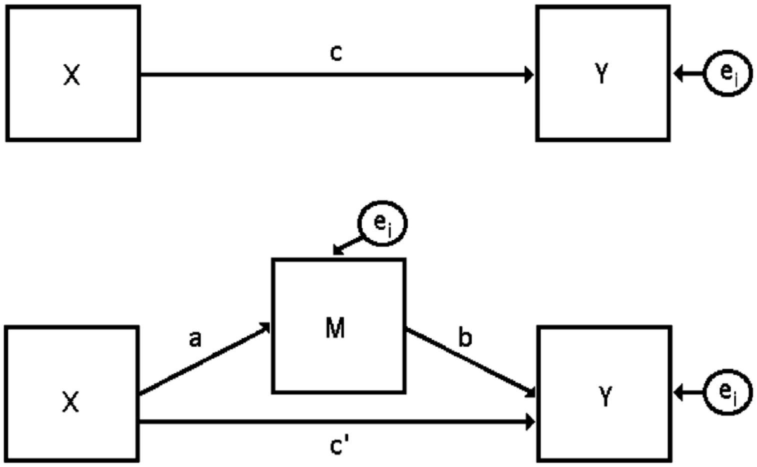

FIGURE 1.
The single mediator model.

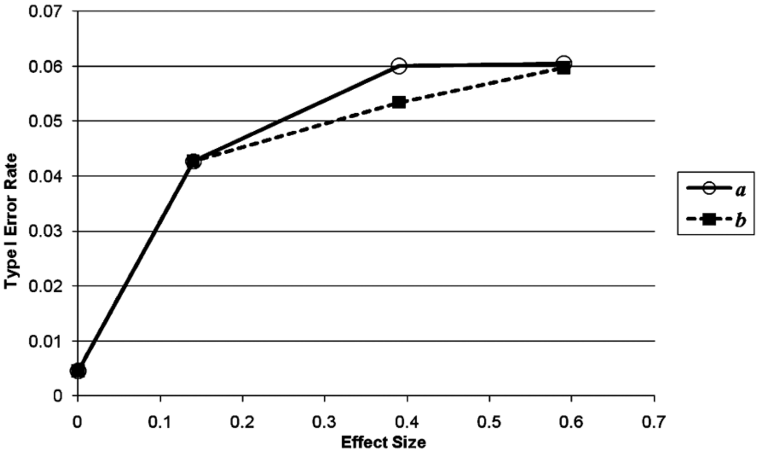

FIGURE 2.
Type I error rate as a function of effect size of *a* and *b*, when the other parameter is equal to zero, collapsing across test and sample size.

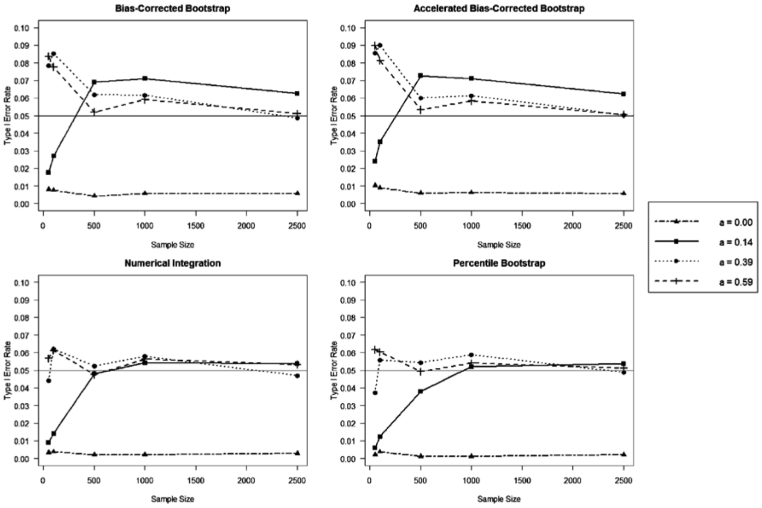

FIGURE 3.

Type I error rate as a function of sample size and effect size of *a*, collapsed across number of bootstrap samples. *Note*. *b* = 0.

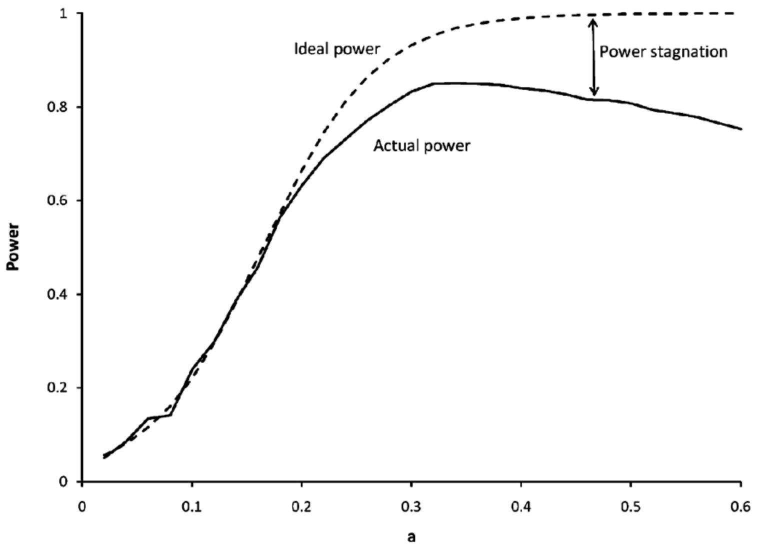

FIGURE 4.

Actual and ideal power curves for the numerical integration test as a function of *a* for condition in which $b = .20$, $c' = .50$, and $n = 160$.

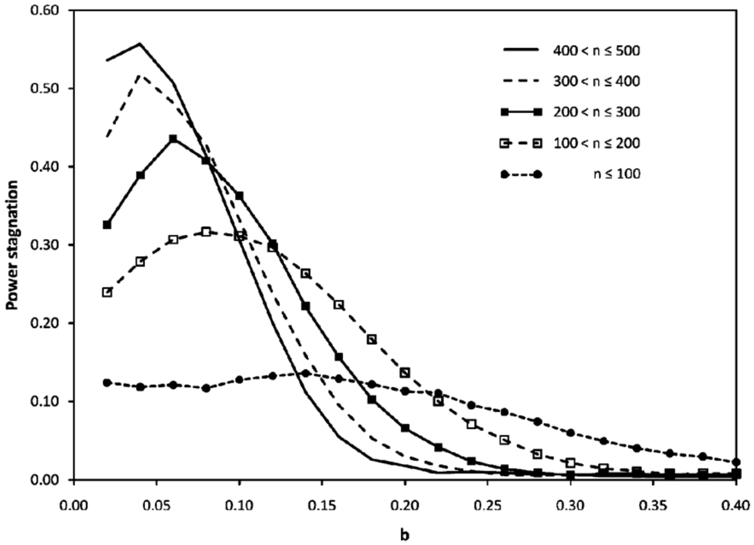

FIGURE 5.
Power stagnation as a function of *b* and sample size. The figure is truncated at *b* = .40 because power stagnation was very near zero for all methods beyond this value.

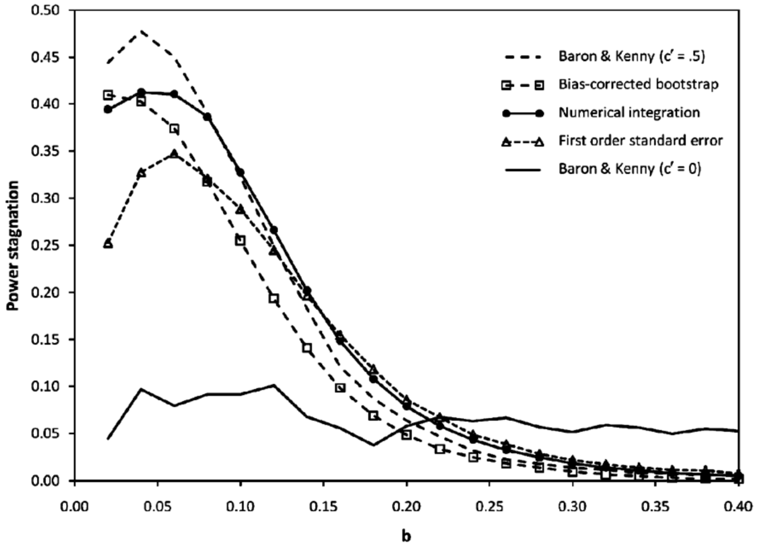

FIGURE 6.

Power stagnation as a function of *b* and test of mediation. The Baron and Kenny (1986) test has results separated by size of c' because it was the only test affected by size of c'. The joint significance test and percentile bootstrap are omitted because their power stagnation values were nearly identical to those of the numerical integration test. The figure is truncated at *b* = .40 because power stagnation was very near zero for all methods beyond this value.

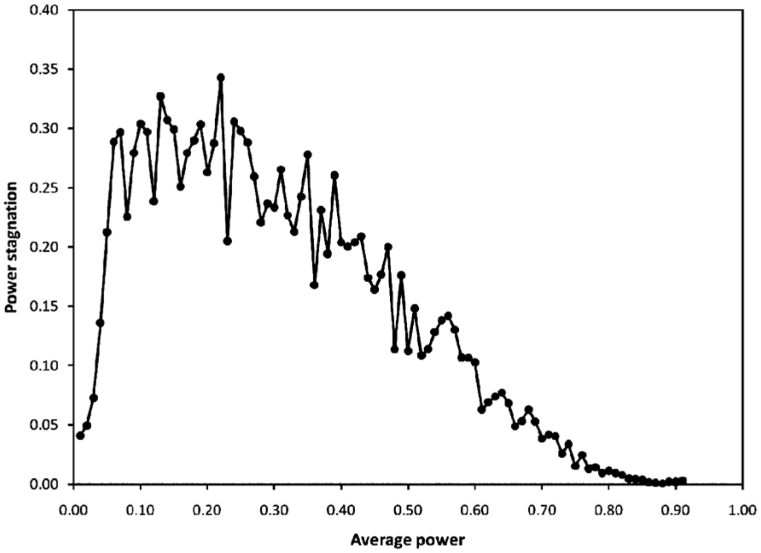

FIGURE 7.

Power stagnation as a function of average power. Average power is the proportion of the area in a power curve figure falling below the power curve, which is also the average level of power across the power curve.

TABLE 1

Results for the Empirical Example for Study 2

	Mediation Model 1		Mediation Model 2	
Value	**Beta (SE)**	**sr^2**	**Beta (SE)**	**sr^2**
a path	0.216 (0.0587) **	0.0466	0.130 (0.0595) *	0.0168
$R^2_{\text{Mult-A}}$		0.0466		0.0168
b path	−0.321 (0.0585) **	0.0984	−0.309 (0.0579) **	0.0940
c' path	0.075 (0.0586)	0.0053	0.038 (0.0577)	0.0014
$R^2_{\text{Mult-BC}'}$		0.0984		0.0940

Note. Beta is a standardized regression coefficient; *SE* is the standard error for the standardized regression coefficient; sr^2 is the squared Type II semipartial correlation; $R^2_{\text{Mult-A}}$ and $R^2_{\text{Mult-BC}'}$ are the squared multiple correlations for the regression models for the *a* path and the *b* and c' paths, respectively.

$^{*}p < .05$.

$^{**}p < .01$.