Methods for Evaluating Mediation Effects: Rationale and Comparison

Zhiyong Zhang and Lijuan Wang
Department of Psychology, University of Virginia, P.O.BOX 400400, Charlottesville, VA 22904-4400, USA

We thank our colleagues at the Center for Developmental and Health Research Methodology at the University of Virginia and the anonymous reviewers for their helpful comments and suggestions.

Abstract

Mediation models have been extensively studied and used in psychological research. In small sample research, a bootstrap resampling method has been shown more effective in testing mediation effects than a large sample method. In the current study, a modified bootstrap method based on resampling residual errors of mediation models is proposed in the framework of path analysis. The proposed method is compared with the large sample method and the bootstrapping raw data method under a variety of conditions. The results show that the bootstrapping error method has better coverage probability and is more efficient than the bootstrapping raw data method when the residual errors are homoscedastic and the effect size is medium to large. In the heteroscedastic case, the bootstrapping raw data method performs best regardless of sample size among the three methods when the effect size is medium to large. However, no methods seem to work well when the effect size is small. A C++ program is provided that implements all the three methods with the modified Brown-Forsythe statistic to test the homoscedasticity of residual errors.

1. Introduction

Mediation models have been studied in psychology for about 80 years. The first psychological research on mediation can be traced back to Woodworth's S-O-R model. Woodrow (1928) found that the active organism intervenes between the stimulus and response are responsible for the effects of stimuli on behavior. Mediation models allow the effect of input variables on output variables to be decomposed into direct and indirect effects. Sometimes, one may find that the relation between two variables is completely due to the existence of a third variable. We then say that the effect of the input variable on the output variable is completely mediated by the third (mediation) variable. Mediation models are also useful for theory development and testing as well as for identification of possible intervention (e.g., Baron and Kenny, 1986; Cole and Maxwell, 2003; Shrout and Bolger, 2002).

The most influential discussion of mediation models was given by Baron and Kenny (1986). Figure 1 depicts the path diagram of a typical mediation model. In this figure, the Y, X, and M represent the dependent or outcome variable, the independent or predictor variable, and the mediation variable, respectively. The e_M and e_Y indicate residuals or measurement errors with variances σ^2_{eM} and σ^2_{eY}. The mediation models can be expressed as two regression equations,

$$\begin{cases} M = a_0 + aX + e_M \\ Y = b_0 + c'X + bM + e_Y \end{cases}, \tag{1}$$

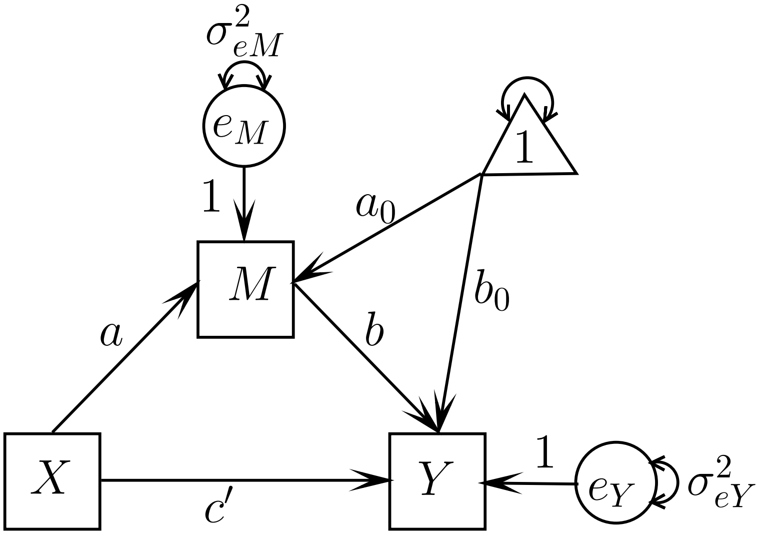

Fig. 1 Path diagram demonstration of a mediation model.

where a_0 and b_0 are intercepts, and a, c' and b are regression coefficients. Thus, the a represents the relation between X and M. The c' represents the relation between X and Y adjusted for M, and the b represents the relation between M and Y adjusted for X. The c' is also called the direct effect of X on Y and ab is called the indirect effect of X on Y through mediation of M. When the mediation effects have occurred, the indirect effect, ab, should be significantly different from 0 (e.g., Shrout and Bolger, 2002).

Statistical approaches to estimating and testing the mediation effects have been discussed extensively in the psychological literature (e.g., Baron and Kenny, 1986; Bollen and Stine, 1990; Shrout and Bolger, 2002; MacKinnon et al., 2002, 2007). Overall, there are two main ways to test the mediation effects. The first one, perhaps also the most influential and widely used one, is the approach outlined by Baron and Kenny (1986). This single sample method (MacKinnon et al., 2002) is based on a large-sample normal approximation test provided by Sobel (1982, 1986) which has low statistical power in many situations (e.g., MacKinnon et al., 2002). The second one may be called the resampling method which is based on the bootstrap resampling procedure (Bollen and Stine, 1990; Efron, 1979, 1987). This method is shown to perform better than the first one in small sample size studies (MacKinnon, Lockwood, and Williams, 2004).

MacKinnon et al. (2002) reviewed and compared 14 methods to test the mediation effects through a Monte Carlo study and found that testing $H_0 : ab = 0$ was the best way to evaluate the mediation effects. MacKinnon, Lockwood, and Williams (2004) also compared the bootstrap resampling method with the single sample method and found that the bootstrap method obtained more accurate confidence limits (See also Shrout and Bolger, 2002). They further suggested that confidence limits if the mediation effects provided much more information than the estimates themselves.

However, several aspects of evaluating the mediation effects have not been examined and need further investigation. First, the coverage probability of the confidence interval for the mediation effects should be investigated (Casella and Berger, 2001). The coverage probability is the probability that a confidence interval can cover the true parameter values (e.g., mediation effects in the current study). For example, if the true mediation effect is .4, for a 95% confidence interval constructed for this mediation effect, it should have a .95 probability of including .4 given that the confidence interval construction method is a good one. If the coverage probability is less than 95%, we will say that the method is too aggressive. Otherwise, the method is too conservative. Note that correct coverage probability is the basis for evaluating power. For example, for the true mediation effect .4, the confidence interval [.1, .35] does not include 0. However, we cannot say that it has power to detect the mediation effect because it does not include the true mediation effect at all. We are not aware of any study discussing the coverage probability for mediation models.

Second, besides the sample size, the normal approximation method can be greatly influenced by the distribution of the residual errors which has not been evaluated yet. If the residual errors are homoscedastic, the normal approximation method will work for the large sample size. However, the normal approximation method may not work well when residual errors are heteroscedastic even for the large sample size data.

Third, the previous bootstrap methods for analyzing the mediation effects were based on resampling of raw data which is usually called the pairwise resampling method (Freedman, 1981) although the original bootstrap method proposed by Efron (1979) was based on resampling of residual errors. A problem of the bootstrapping raw data method is that it could lead to large standard errors of parameter estimates due to the change of the design matrix when conducting resampling. For the purpose of illustration, consider a simple regression model $y = bx + e$. With the least squares method, an estimator of b is $\hat{b} = cov(y, x)/var(x)$. For the data set with small sample size and/or outliers, the variance of x may become very unstable in bootstrap samples. For an extreme case, if one is unlucky to resample a data set with all the same x values, the variance of x will become 0 and the estimate of b will become infinity. A more realistic case is that some resampled samples may include the outlier and others may not. Then the variance of x may fluctuate a lot which will result in the rejection of the true mediation effects.

The purpose of this article is to evaluate and compare different methods for analyzing mediation effects under a variety of conditions. First, we review the normal approximation data method and the bootstrapping raw data method. Second, we present the procedure of the bootstrap method based on the resampling of residual errors of mediation models. Third, we conduct a simulation study to evaluate the performance of the three methods with different sample sizes, effect sizes, and distributions of residual errors. Finally, we give suggestions on how to choose different methods in the empirical data analysis.

2. Methods for Evaluating Mediation Effects

In this section, we will present the three mediation analysis methods . For the single sample method, MacKinnon et al. (2002) have reviewed and compared 14 methods. Here, we only focus on one of the methods, the normal approximation method, which is based on the delta method (Casella and Berger, 2001). For the resampling method, we will present the implementation procedure for bootstrapping raw data and residual errors methods.

2.1. Normal Approximation Method

The normal approximation method is the most influential and widely used mediation analysis method. This method assumes that the mediation effect has a normal distribution which can be constructed through the delta method. The distribution of ab can be estimated by $N(\hat{a}\hat{b}, \hat{s}_{ab})$ with $\hat{a}$ and $\hat{b}$ estimated from the mediation model. The standard error of the mediation effect, $\hat{s}_{ab}$, can be calculated approximately by the delta method based on the first order Taylor expansion given the variances of a and b obtained from the path model in Eq (1) (Casella and Berger, 2001). Specifically, the standard error of ab is approximated by $\hat{s}_{ab} = \sqrt{\hat{b}^2\hat{\sigma_a^2} + \hat{a}^2\hat{\sigma_b^2}}$ with the estimated variances $\hat{\sigma_a^2}$ and $\hat{\sigma_b^2}$ for $\hat{a}$ and $\hat{b}$. The more accurate standard errors can be constructed by including the covariance and higher order terms of the Taylor expansion. Based on this standard error, we can construct a confidence interval for ab as

$$[\hat{a}\hat{b} - z_{1-\alpha/2}\hat{s}_{ab}, \hat{a}\hat{b} + z_{1-\alpha/2}\hat{s}_{ab}],$$

where z_α represents the $100\alpha\%$ percentile of the standard normal distribution.

Many researchers have pointed out the disadvantages of this single sample method. First, because the estimate of the standard error of the mediation effect is based on asymptotic distribution, it usually requires a large sample size (e.g., Shrout and Bolger, 2002). Second, the constructed confidence interval is symmetric. However, the distribution of ab is usually nonnormal and the confidence interval is typically asymmetric (Bollen and Stine, 1990; MacKinnon et al., 2002). Ignoring these problems will reduce the power to detect the mediation effects (Shrout and Bolger, 2002).

2.2. Bootstrapping Raw Data Method

The resampling method is usually referred as to the bootstrap method (Efron, 1979, 1987). It was first employed in mediation analysis by Bollen and Stine (1990) and has been studied in a variety of research contexts (e.g., MacKinnon, Lockwood, and Williams, 2004; Mallinckrodt et al., 2006; Preacher and Hayes, 2004; Shrout and Bolger, 2002). This method has no distributional assumption on the indirect effect ab. Instead, it approximates the distribution of ab using its bootstrap distribution. The bootstrap method was shown to be more appropriate for studies with the sample size of 20-80 than the single sample method (Shrout and Bolger, 2002).

Currently, the bootstrap method of the mediation effects generally follow the bootstrapping raw data procedure as used in Bollen and Stine (1990). This procedure can be summarized as follows.

1. Using the *original data set* (Sample size $= N$) as a population, draw a bootstrap sample of N persons with paired Y, X, and M randomly with replacement from the original data set.

2. From the bootstrap sample, estimate $\hat{a}\hat{b}$ through the ordinary least squares (OLS) method based on a set of regression models.

3. Repeat Steps 1 and 2 for a total of B times. The B is called the bootstrap sample size.

4. The empirical distribution of $\hat{a}\hat{b}$ based on this bootstrap procedure can be viewed as the distribution of ab. The $(1-\alpha)\times 100\%$ confidence interval of ab can be constructed using the $(\alpha/2)\times 100\%$ and $(1-\alpha/2)\times 100\%$ percentile of the empirical distribution.

Besides the confidence interval above, the other bootstrap confidence intervals can also be constructed (DiCiccio and Efron, 1996; Hall, 1988; MacKinnon, Lockwood, and Williams, 2004).

2.3. Bootstrapping Error Method

Although the original bootstrap method proposed by Efron (1979) is based on resampling of independently and identically distributed data, the method used in mediation analysis is generally based on the resampling of paired raw data (Freedman, 1981). Here, we first outline the procedure to implement the bootstrapping error method and then compare it with the bootstrapping raw data method. For the bootstrapping error method, the empirical distribution of ab can be constructed by the following steps:

1. Estimate the parameters $\hat{a}_0, \hat{a}, \hat{c}', \hat{b}_0$, and $\hat{b}$ by using SEM/Path analysis method.
2. Estimate the residual errors $\hat{e}_M$ and $\hat{e}_Y$ by plugging $\hat{a}_0, \hat{a}, \hat{c}', \hat{b}_0$, and $\hat{b}$ into $\hat{e}_M = M - \hat{a}_0 - \hat{a}X$ and $\hat{e}_Y = Y - \hat{b}_0 - \hat{c}'X - \hat{b}M$. Pair the residual errors to be $(\hat{e}_M, \hat{e}_Y)$.
3. Using the *estimated residual errors* (Sample size = N) as a population, draw a bootstrap sample of N persons with paired residual errors $(\hat{e}_{Mb}, \hat{e}_{Yb})$ randomly with replacement.
4. Calculate $\hat{Y}_b, \hat{M}_b$ from X and $(\hat{e}_{Mb}, \hat{e}_{Yb})$ with regression parameters $\hat{a}_0, \hat{a}, \hat{c}', \hat{b}_0$, and $\hat{b}$ by using $\hat{M}_b = \hat{a}_0 + \hat{a}X + \hat{e}_{Mb}$ and $\hat{Y}_b = \hat{b}_0 + \hat{c}'X + \hat{b}\hat{M}_b + \hat{e}_{Yb}$ to form a new bootstrap sample $(\hat{Y}_b, \hat{M}_b, X)$. Note that X is the same for all bootstrap samples.
5. Estimate the regression parameters using Step 1 and calculate $\hat{a}\hat{b}$.
6. Repeat Steps 3-5 for a total of B (Bootstrap sample size) times.
7. The empirical distribution of $\hat{a}\hat{b}$ based on this bootstrap procedure can be viewed as the distribution of ab. The $(1-\alpha) \times 100\%$ confidence interval of ab can be constructed using the $(\alpha/2) \times 100\%$ and $(1-\alpha/2) \times 100\%$ percentile of the empirical distribution.

There are two major differences between the bootstrapping raw data method and the bootstrapping error method. First, the design matrix (X) does not change over the bootstrap samples in the bootstrapping error method. However, in the bootstrapping raw data method, the design matrix generally changes over the bootstrap samples. Second, the bootstrapping error method requires the residual errors to be independently and identically distributed. However, the bootstrapping raw data method only requires that data to be sampled are independently distributed. For example, the distribution of Y is only independently but not identically distributed.

3. Simulation Study

To evaluate and compare the performance of the three methods discussed above, a systematic simulation study is designed and implemented.

3.1. Simulation Design

MacKinnon and colleagues have designed simulation studies for comparing different methods on evaluating the mediation effects (e.g., MacKinnon et al., 2002; MacKinnon, Lockwood, and Williams, 2004). The current simulation design was built on their simulation study designs. The following factors were considered in the current simulation.

Sample size. Three sample sizes were used in the simulation study, N =25, 50, and 100.

Effect size. Population parameter values were chosen to represent the effect size of 0, small, medium, and large (Cohen, 1988; MacKinnon et al., 2002). In detail, four combinations of a and b were considered: $a = b = 0$, $a = b = .14$, $a = b = .39$, and $a = b = .59$.

Distribution of residual errors. Two kinds of distributions for residual errors were chosen. In the first case, residuals errors were independently, identically and normally distributed as $e_{Mi} \sim N(0, 1), i = 1, \ldots,$ N and $e_{Yi} \sim N(0, 1), i = 1, \ldots,$ N. In the second case, residuals errors were given independent but not identical normal distribution as $e_{Mi} \sim N(0, X_i), i = 1, \ldots,$ N and $e_{Yi} \sim N(0, X_i), i = 1, \ldots,$ N.

In total, 24 conditions were considered in the simulation. For each condition, R=10000 sets of data were simulated and the 80% and 95% confidence intervals for $\hat{a}\hat{b}$ were constructed. In each condition, we focus on the confidence intervals to compare the results.

Coverage probability. For an ideal confidence interval, its coverage probability should match its confidence level. For example, a 95% confidence interval should have a .95 coverage probability. Only with appropriate coverage probability, comparisons of power and confidence intervals are meaningful.

Power. Power is the probability that a test will reject a false null hypothesis. A better confidence interval should have a larger power.

Confidence intervals. We focus on the average confidence limits and their standard deviations. The smaller the standard deviations, the more efficient the confidence limits.

3.2. Simulation Results

Now we report the results from the simulation study. Because of the space limitation, we only provide the results from the conditions $a = b = .14$ and $a = b = .59$ for both independently and identically distributed residual error (iid) and independently but not identically distributed residual error (non-iid) cases.[*]

The results for the iid case are given in Tables 1. and 2. From Table 1. ($a = b = .14$), we can conclude when the effect size was small, (1) no method worked well with a small sample size ($N = 25$) based on coverage probabilities and all confidence intervals were overestimated; (2) with $N = 50$, the normal approximation method seemed to work best although its power was lowest among all three methods; (3) with $N = 100$, the bootstrapping error method and the bootstrapping raw data method worked equally well.

When the effect size is large ($a = b = .59$, Table 2.), the bootstrapping error method achieved the most accurate coverage probability and largest power from the small sample size ($N = 25$) to the large sample size ($N = 100$). When $N = 100$, there was no significant difference in confidence limits and power between the bootstrapping error method and bootstrapping raw data method. The same conclusions were reached for the medium effect size ($a = b = .39$) situation.

The results for the non-iid residual errors case are given in Tables 3. and 4. When the effect size is small ($a = b = .14$, Table 3.), (1) the normal approximation method had the best coverage probability and the largest power with the small sample size ($N = 25$); (2) the bootstrapping error method seemed to work best based on coverage probability with $N = 50$ and $N = 100$; (3) the normal approximation method gave the largest power in all sample sizes because it had the smallest confidence intervals and thus the underestimated coverage probabilities.

[*] Interested readers can view the full results at http://medci.psychstat.org.

Table 1. Small effect size ($a = b = .14$), iid residuals errors.

		Sample Size					
		25		50		100	
		80%	95%	80%	95%	80%	95%
CP	Norm	0.9187	0.9849	0.8035	0.9536	0.7745	0.9085
	Error	0.9356	0.9932	0.8612	0.9856	0.7949	0.9683
	Raw	0.9165	0.9889	0.8369	0.9836	0.7913	0.9601
Power	Norm	0.0403	0.002	0.0803	0.0063	0.1829	0.0169
	Error	0.0565	0.0095	0.1189	0.0239	0.2467	0.0645
	Raw	0.0631	0.011	0.1264	0.0261	0.2522	0.0698
CI	True	[-38,94]	[-91,174]	[-14,66]	[-39,109]	[-3,50]	[-14,75]
	Norm	[-64,107]	[-11,152]	[-28,68]	[-53,93]	[-1,49]	[-25,64]
	s.d.	62,92	75,112	3,54	33,64	16,33	15,38
	Error	[-72,121]	[-145,204]	[-28,73]	[-62,117]	[-8,52]	[-25,77]
	s.d.	69,96	93,125	33,55	42,69	17,33	21,4
	Raw	[-72,119]	[-148,204]	[-28,73]	[-62,116]	[-8,51]	[-25,76]
	s.d.	72,97	102,131	33,56	44,71	17,33	21,41

Note. CP: coverage probability; CI: confidence interval; Norm: normal approximation method; Error: bootstrapping error method; Raw: bootstrapping raw data method; s.d.: standard deviation of confidence limits; CI was rescaled by multiplying 1000.

When the effect size is large ($a = b = .59$, Table 4.), (1) the coverage probabilities from all methods were smaller than they should be and all the confidence intervals were smaller than the true confidence intervals; (2) the bootstrapping raw data method seemed to give the closest coverage probabilities but the power was the lowest among the three methods; (3) the normal approximation method had the largest power although the coverage probabilities were the most underestimated ones. The results were similar when the effect size was medium ($a = b = .39$).

4. Conclusions and Discussion

In this study, we investigated and compared the three methods for evaluating the mediation effects. Results from the simulation study support the following conclusions. First and most important, when effect size was medium or large and the residual errors were iid, the bootstrapping error methods had the best coverage probability and the largest power. When the effect size was medium or large and the residual errors were non-iid, the bootstrapping raw data methods had the best coverage probability. Second, when the effect size was small, we did not find any consistent conclusions for different methods. Third, if researchers only focus on the power of different methods, the conclusions may be very misleading. For example, the normal approximation method had the largest power when the residuals errors were non-iid. However, this method cannot be used at all in this situation because the constructed confidence intervals were not correct.

To help researchers implement the methods discussed in this study, a C++ program, MedCI, was customized and provided (Zhang and Wang, 2007). This program can estimate the mediation effect from the mediation model from all three methods automatically. For empirical research, testing the homoscedasticity of the residual errors became especially important because it was directly related to the choice of different methods. Considering that the sample size may be very small, we employed the modified Brown-Forsythe test in our study (Brown and

Table 2. Large effect size ($a = b = .59$), iid residuals errors.

		Sample Size					
		25		50		100	
		80%	95%	80%	95%	80%	95%
CP	Norm	0.7781	0.9169	0.7854	0.9304	0.7895	0.9373
	Error	0.7969	0.9449	0.7946	0.9459	0.7979	0.9453
	Raw	0.7821	0.9334	0.782	0.9394	0.7886	0.9407
Power	Norm	0.8301	0.4614	0.9871	0.9124	1	0.9995
	Error	0.8419	0.5785	0.9891	0.9425	1	0.9996
	Raw	0.8411	0.5625	0.9888	0.9387	1	0.9996
CI	True	[130,591]	[48,769]	[197,511]	[136,616]	[242,460]	[195,529]
	Norm	[12,577]	[-1,698]	[193,506]	[11,589]	[239,456]	[181,513]
	s.d.	14,236	124,265	1,15	89,165	73,99	67,106
	Error	[122,599]	[21,771]	[196,515]	[13,621]	[241,46]	[193,528]
	s.d.	141,241	14,278	101,152	92,168	74,99	68,107
	Raw	[125,594]	[17,769]	[198,513]	[131,618]	[241,459]	[194,527]
	s.d.	144,245	146,291	102,153	94,172	74,99	69,108

Note. Same as the above table.

Forsythe, 1974; Mehrotra, 1997; Rubin, 1983). This test has been incorporated into the provided program.

There are a couple of perspectives of testing the mediation effects which can be investigated in the future. First, with a small effect size, no methods seemed to work well especially when the sample size was also small. Furthermore, for the non-iid case, the normal approximation method and the bootstrapping error methods were surprisingly better than the bootstrapping raw data method. Thus, the three methods under the small effect size and the small sample size conditions should be further investigated. Second, when the residuals errors were non-iid, the coverage probabilities were all underestimated for the bootstrapping raw data method. Different confidence intervals, such as the studentized confidence interval and the bias-corrected confidence interval, can be employed and evaluated to obtain better coverage probabilities (DiCiccio and Efron, 1996).

References

Baron, R. M., & 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.

Bollen, K. A., & Stine, R. A. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology, 20*, 115–140.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16*, 129–132.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*, 558–577.

Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). Duxbury Press.

DiCiccio T. J., & Efron B. (1996). Bootstrap confidence intervals (with discussion). *Statistical Science, 11*, 189–228.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics, 7*(1), 1–26.

Table 3. Small effect size ($a = b = .14$), non-iid residuals errors.

		Sample Size					
		25		50		100	
		80%	95%	80%	95%	80%	95%
CP	Norm	0.7856	0.9474	0.6594	0.9193	0.5771	0.8101
	Error	0.8472	0.9634	0.8109	0.9583	0.7113	0.9427
	Raw	0.902	0.9888	0.8871	0.9874	0.8462	0.9848
Power	Norm	0.1165	0.0323	0.1579	0.0441	0.2389	0.0772
	Error	0.0979	0.0286	0.1242	0.0395	0.1819	0.0622
	Raw	0.055	0.0094	0.0747	0.0121	0.1042	0.0201
CI	True	[-105,161]	[-237,326]	[-51,107]	[-117,203]	[-22,75]	[-55,132]
	Norm	[-92,133]	[-151,192]	[-42,82]	[-74,115]	[-16,55]	[-35,74]
	s.d.	135,159	148,181	7,92	74,104	38,56	38,63
	Error	[-12,165]	[-221,272]	[-59,102]	[-113,163]	[-26,68]	[-55,103]
	s.d.	147,174	186,218	77,101	96,125	41,61	5,75
	Raw	[-153,199]	[-28,338]	[-78,123]	[-148,203]	[-38,82]	[-77,13]
	s.d.	16,19	218,253	83,11	113,143	44,68	59,88

Note. Same as the above table.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171–185.

Freedman, D. A. (1981). Bootstrapping regression models, *Annals of Statististics, 9*, 1218–1228.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statististics, 16*, 927–985.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614.

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

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*(1), 99–128.

Mallinckrodt, B., Abraham, T. W., Wei, M., & Russell, D. W. (2006). Advance in testing statistical significance of mediation effects. *Journal of Counseling Psychology, 53*, 372–378.

Mehrotra, D. V. (1997). Improving the Brown-Forsythe solution to the generalized Behrens- Fisher problem. *Communications in Statistics-Simulation and Computation, 26*, 1139– 1145.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717–731.

Rubin, A. S. (1983). The use of weighted contrast in analysis of models with heterogeneity of variance. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, pp. 347–352.

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

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology.* San Francisco: Jossey-Bass, pp. 290–312.

Table 4. Large effect size ($a = b = .59$), non-iid residuals errors.

		Sample Size					
		25		50		100	
		80%	95%	80%	95%	80%	95%
CP	Norm	0.5553	0.7362	0.5409	0.735	0.5544	0.7484
	Error	0.6451	0.8309	0.6454	0.8387	0.6689	0.859
	Raw	0.7505	0.9134	0.7499	0.9179	0.7749	0.9311
Power	Norm	0.6833	0.4381	0.8939	0.7512	0.9906	0.9661
	Error	0.6389	0.3885	0.8571	0.6795	0.9804	0.9276
	Raw	0.5423	0.2616	0.7955	0.5527	0.9649	0.8691
CI	True	[19,758]	[-108,1075]	[105,623]	[18,824]	[174,541]	[103,671]
	Norm	[117,58]	[-5,703]	[191,508]	[108,592]	[238,457]	[18,514]
	s.d.	253,375	234,415	177,251	16,273	126,165	116,176
	Error	[8,643]	[-62,84]	[158,556]	[68,687]	[209,493]	[146,581]
	s.d.	249,397	264,462	17,263	166,297	121,173	114,19
	Raw	[23,708]	[-139,963]	[118,604]	[016,77]	[18,528]	[109,642]
	s.d.	232,406	262,483	158,274	159,316	114,18	107,203

Note. Same as the above table.

Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. In N. Tuma (Ed.), *Sociological methodology*. Washington, DC: American Sociological Association, pp. 159–186.

Woodworth, R. S. (1928). Dynamic psychology. In C. Murchison (Ed.), *Psychologies of 1925*. Worcester, MA: Clark University Press, pp. 111–126.

Zhang, Z., & Wang, L. (2007). MedCI: Mediation Confidence Intervals, Version 3.0. Retrievable from http://medci.psychstat.org