Organizational
Research Methods
Volume XX Number X
Month XXXX xx-xx

10.1177/1094428107300344
http://orm.sagepub.com
hosted at
http://online.sagepub.com

Tests of the Three-Path Mediated Effect

Aaron B. Taylor
David P. MacKinnon
Jenn-Yun Tein
Arizona State University

In a three-path mediational model, two mediators intervene in a series between an independent and a dependent variable. Methods of testing for mediation in such a model are generalized from the more often used single-mediator model. Six such methods are introduced and compared in a Monte Carlo study in terms of their Type I error, power, and coverage. Based on its results, the joint significance test is preferred when only a hypothesis test is of interest. The percentile bootstrap and bias-corrected bootstrap are preferred when a confidence interval on the mediated effect is desired, with the latter having more power but also slightly inflated Type I error in some conditions.

Keywords: *mediation; bootstrapping*

Mediational effects are commonly studied in organizational behavior research. For example, Stewart and Barrick (2000) studied the relation between self-rated interdependence in production teams and the teams' supervisor-rated performance. They found that a set of intrateam process variables—communication, conflict, shirking, and flexibility—mediated the relation between the teams' interdependence and performance. In another study, employees' perceived control of their time was found to partially mediate the relations between the independent variables of workload, job autonomy, and planning behavior and the dependent variables of work strain, job satisfaction, and job performance (Claessens, Van Eerde, Rutte, & Roe, 2004). Among other effects, workload reduced perceived control of time and perceived control of time improved job satisfaction.

Mediational models are also common in other social sciences. A well-known example of mediation in psychology is that intentions mediate the effect of attitude on behavior (Ajzen & Fishbein, 1980). In sociology, son's educational achievement is thought to mediate the effect of father's socioeconomic status on son's socioeconomic status (Duncan, Featherman, & Duncan, 1972). In experimental prevention research, a prevention program is designed to change social norms regarding smoking, which are assumed to be causally related to smoking (e.g., MacKinnon & Dwyer, 1993).

In all of the above examples, one mediator transmits the influence of an independent variable to a dependent variable. Some theories are based on a long mediation chain even though analyses often focus on single mediators. Cook and Campbell (1979) called this chain of effects the micromediational chain. Analyses of micromediational chains longer than two paths (one mediator) are also becoming common. For example, Tekleab, Bartol,

Authors' Note: This research was supported by U.S. Public Health Service Grant DA09757 to David P. MacKinnon. We thank three anonymous reviewers and the guest editor for their helpful comments on an earlier version of this article.

and Liu (2005) found support for a model in which the effect of pay on turnover was mediated by two variables acting in turn, perceived distributive justice (fairness of allocation) and pay raise satisfaction. Employees who were paid more perceived higher levels of distributive justice, which led to higher reported levels of satisfaction with pay raises, which in turn reduced the probability they would voluntarily leave the company. In another study of employee turnover, Allen and Griffeth (2001) found that job performance positively affected employees' perceived employment alternatives, which positively affected their intention to leave (turnover intention), which in turn affected actual turnover.

Such longer chain mediational models can also be found in other social sciences. For example, a study of divorced mothers tested the hypothesis that the effect of negative life events on parenting behaviors would be mediated by two variables in turn: psychological distress and avoidant coping (Tein, Sandler, & Zautra, 2000). In another study of an intervention designed to increase mammography screening, the effect was mediated by perceived susceptibility to breast cancer and perceived benefits of screening (Aiken, Gerend, & Jackson, 2001). Going beyond two mediator series, McGuire (1980, pp. 102-103) proposed a model in which a series of seven variables mediates the effects of exposure to health communication on behavior.

Although McGuire's (1980) model is an extreme example, there are clearly models being proposed that include more than a single mediator in the causal chain between independent and dependent variables. The single-mediator case has been extensively studied. A number of methods of testing for mediation in this context have been proposed (e.g., Baron & Kenny, 1986; MacKinnon, Fritz, Williams, & Lockwood, in press; Sobel, 1982), and their performance has been discussed and compared (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; MacKinnon, Lockwood, & Williams, 2004; Shrout & Bolger, 2002). Methods of testing for single-mediator effects have not yet been generalized to testing for longer mediational chains, though. The purpose of this article is to extend several methods used in the two-path (single-mediator) context to the three-path (two mediators in series) context, with the ultimate aim of concluding which methods might perform best in testing mediational chains of any length.

Defining the Mediated Effect

The three-path mediation model is depicted as a path diagram in Figure 1. Estimating the model requires that the following three regression equations be estimated:

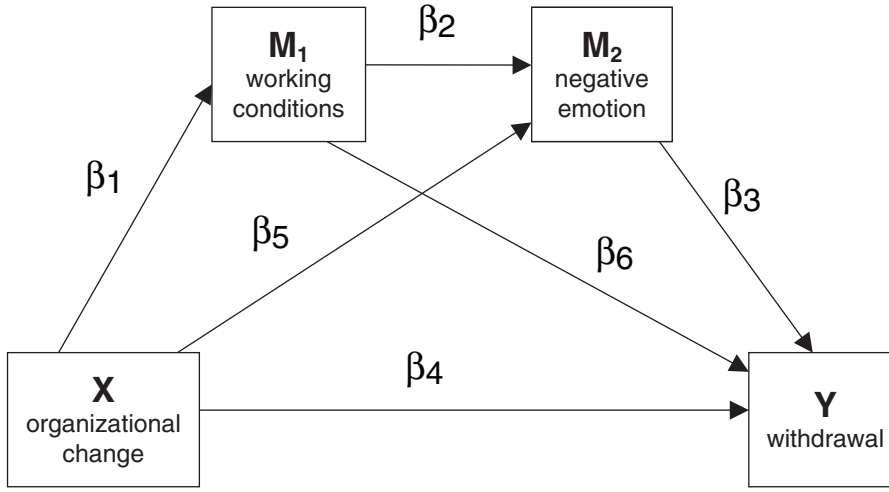
$$M_1 = \beta_{01} + \beta_1 X + \varepsilon_1, \tag{1}$$

$$M_2 = \beta_{02} + \beta_2 M_1 + \beta_5 X + \varepsilon_2, \tag{2}$$

$$Y = \beta_{03} + \beta_4 X + \beta_3 M_2 + \beta_6 M_1 + \varepsilon_3. \tag{3}$$

In these equations, Y is the dependent variable, X is the independent variable, and M_1 and M_2 are the two mediators. In the first equation, β_1 is the regression of M_1 on X. In the second equation, β_2 is the regression of M_2 on M_1, and β_5 is the regression of M_2 on X. In the third equation, β_3 is the regression of Y on M_2, β_4 is the regression of Y on X, and β_6 is

Figure 1
Path Diagram of the Three-Path Mediated Effect Model



Note: Residuals are omitted. Variable labels refer to data based on Kiefer (2005).

the regression of Y on M_1. The intercepts in the equations are β_{01}, β_{02}, and β_{03}. The residuals are ε_1, ε_2, and ε_3.

There are a number of different effects of X on Y that might be defined using this model. The direct effect of X on Y, controlling for both mediators, is β_4 in Equation 3. Mediated effects are estimated by the product of the coefficients for each of the paths in the mediational chain (Alwin & Hauser, 1975). Therefore, the total mediated effect of X on Y, the effect passing through either mediator, is $\beta_1\beta_2\beta_3 + \beta_1\beta_6 + \beta_5\beta_3$. This effect can be broken down into the three-path mediated effect, which is the effect passing through both mediators ($\beta_1\beta_2\beta_3$), and the two-path mediated effects, the effects passing through only one of the mediators ($\beta_1\beta_6$ and $\beta_5\beta_3$). Another effect that may be studied is the mediated effect passing through one mediator, such as $\beta_1\beta_2\beta_3 + \beta_5\beta_3$ for M_2, for which Sobel (1982) introduced a standard error formula. This article focuses solely on methods of testing the three-path mediated effect as this is the effect most likely to be of interest to researchers (James, Mulaik, & Brett, 2006).

Another equation is sometimes considered in testing for mediation in addition to Equations 1-3. The most commonly used method of testing for mediation in the single-mediator context, the approach of Kenny and colleagues (Baron & Kenny, 1986; Judd & Kenny, 1981; Kenny, Kashy, & Bolger, 1998), requires that the total effect of the independent variable on the dependent variable also be tested:

$$Y = \beta_{00} + \tau X + \varepsilon_0. \tag{4}$$

This test is considered necessary because if the total effect τ is significantly nonzero, this "establishes that there is an effect that may be mediated" (Kenny et al., 1998, p. 259), whereas if τ is nonsignificant, there is no effect of X on Y to be mediated.

In this article, we take the position that the test in Equation 4 is not necessary to establish mediation. There are two reasons for continuing to test for mediation even in the presence of a nonsignificant total *X* to *Y* relation. First, as noted by Shrout and Bolger (2002, p. 429), mediational analysis can provide a more powerful test of the relation between the independent and dependent variables than the simple regression in Equation 4 provides. This means that the test of the total effect might be nonsignificant simply because it is not as powerful as the test of mediation. On the basis of this possibility, Shrout and Bolger recommended dropping the test of the total relation under circumstances where the measurement of the dependent variable is far removed in time from the measurement of the independent variable.

The second reason for testing for mediation even if τ is nonsignificant is that suppression (inconsistent mediation) may occur, meaning that the mediated and direct effects have opposite signs. As noted by MacKinnon et al. (2002, p. 87) for the single-mediator model, if the mediated and direct effects have opposite signs, the total effect may be near zero, even though the mediated effect is significantly nonzero. The three-path situation is more complex, as the signs of the two-path mediated effects $\beta_1\beta_6$ and $\beta_5\beta_3$ may also be either positive or negative, but the point remains the same. The total effect may not differ significantly from zero even in the presence of significant mediation because of the different signs of the effects that make it up.

Methods of Testing for Mediation

Following the framework proposed by MacKinnon et al. (2002; MacKinnon et al., 2004), methods of testing for mediation can be put into four categories. These are causal-steps tests, product-of-coefficients tests, difference-in-coefficients tests, and resampling methods.

Causal-steps tests. The approach of Kenny and colleagues (Baron & Kenny, 1986; Judd & Kenny, 1981; Kenny et al., 1998) mentioned above is classified as a causal-steps test. Although it was proposed for the single-mediator situation, it can be extended to the three-path model. This approach requires the test of the total effect of *X* on *Y* to be significant, though, so it is not applied in this article.

Another causal-steps test, called the joint significance test by MacKinnon et al. (2002), is based on the definition of mediation offered by James and Brett (1984). It differs from the Kenny et al. approach in that it does not require the overall relation between the predictor and the outcome to be significant. Although it was proposed for the single-mediator situation, it can be easily generalized to the three-path context. In a three-path mediational model, the joint significance test finds evidence for mediation if each of the three paths in the mediated effect is significantly nonzero (see Table 1). These paths are β_1, β_2, and β_3 in Figure 1 and Equations 1-3. The corresponding sample coefficients are b_1, b_2, and b_3. As noted by MacKinnon et al., the major weaknesses of the joint significance test are that it does not provide an estimate of the mediated effect and that it cannot be easily used to construct a confidence interval. Its notable strengths are its simplicity—it merely requires null hypothesis tests for three regression coefficients—and its control of Type I error. Because all three paths must be significant for the mediated effect to be significant,

Table 1
Methods of Testing the Three-Path Mediated Effect

Method	Test ($\alpha = .05$)
Joint significance	Reject Hypothesis$_0$ if $\lvert b_1\rvert/s_{b1} > t_{.975(n-2)}$ and $\lvert b_2\rvert/s_{b2} > t_{.975(n-3)}$ and $\lvert b_3\rvert/s_{b3}\ t_{.975(n-4)}$, where $t_{.975(df)}$ is the critical t value for a two-tailed test given the *df*. For example, $t_{.975(100)} = 1.98$.
Multivariate delta standard error	Reject Hypothesis$_0$ if 95% confidence interval $= b_1b_2b_3 \pm z_{.975}(s^2_{\text{multivariate delta}})^{1/2}$ does not include zero, where $s^2_{\text{multivariate delta}} = b_1^2b_2^2s_{b3}^2 + b_1^2b_3^2s_{b2}^2 + b_2^2b_3^2s_{b1}^2$ and $z_{.975} = 1.96$.
Unbiased standard error	Reject Hypothesis$_0$ if 95% confidence interval $= b_1b_2b_3 \pm z_{.975}(s^2_{\text{unbiased}})^{1/2}$ does not include zero, where $s^2_{\text{unbiased}} = b_1^2b_2^2s_{b3}^2 + b_1^2b_3^2s_{b2}^2 + b_2^2b_3^2s_{b1}^2 - b_1^2s_{b2}^2s_{b3}^2 - b_2^2s_{b1}^2s_{b3}^2 - b_3^2s_{b1}^2s_{b2}^2 + s_{b1}^2s_{b2}^2s_{b3}^2$.
Exact standard error	Reject Hypothesis$_0$ if 95% confidence interval $= b_1b_2b_3 \pm z_{.975}(s^2_{exact})^{1/2}$ does not include zero, where $s^2_{\text{exact}} = b_1^2b_2^2s_{b3}^2 + b_1^2b_3^2s_{b2}^2 + b_2^2b_3^2s_{b1}^2 + b_1^2s_{b2}^2s_{b3}^2 + b_2^2s_{b1}^2s_{b3}^2 + b_3^2s_{b1}^2s_{b2}^2 + {}^2_{b1}s_{b2}^2s_{b3}^2$.
Percentile bootstrap	Draw a large number of bootstrap samples and estimate $b_1b_2b_3$ in each to form bootstrap distribution. Endpoints of a 95% confidence interval are 2.5th and 97.5th percentiles of distribution. Reject Hypothesis$_0$ if confidence interval does not include zero.
Bias-corrected bootstrap	Form bootstrap distribution as above. Find p, proportion of the distribution greater than original sample $b_1b_2b_3$. Calculate $z_{\text{lower}} = -1.96 + 2z_0$ and $z_{\text{upper}} = 1.96 + 2z_0$, where z_0 is the z score corresponding to probability p. For example, for $p = .55$, $z_0 = 0.13$. End points of a 95% confidence interval are percentile ranks from the bootstrap distribution corresponding to normal percentiles for z_{lower} and z_{upper}. Reject Hypothesis$_0$ if confidence interval does not include zero.

its null hypothesis rejection rate is the product of the probabilities of rejecting the individual coefficients' null hypotheses. Therefore, its Type I error should not exceed the nominal level of .05 even if two of the paths are so large that their null hypotheses are rejected with probability 1: $.05 \times 1 \times 1 = .05$. Consistent with this expectation, MacKinnon et al. found that the joint significance test controlled Type I error well and had good power.

Product-of-coefficients tests. Another class of methods of testing for mediation are what MacKinnon et al. (2002) called product-of-coefficients tests. As noted above, $\beta_1\beta_2\beta_3$ is the mediated effect; its sample estimator is $b_1b_2b_3$. Generalizing from the single-mediator context, a product-of-coefficients test of the mediated effect divides $b_1b_2b_3$ by its estimated standard error and refers the result to a standard normal distribution (although the distribution of $b_1b_2b_3$ is likely to not be normal). The standard error may be estimated using different approaches. The derivation of three estimators of the variance (and therefore the standard error) of $b_1b_2b_3$ is shown in the appendix. The estimators are the multivariate delta estimator, the unbiased estimator, and the exact estimator. Their formulas are shown in Table 1. The first extends the work of Sobel (1982) from the two-path to the three-path situation, and the other two are based on Goodman's (1960) work. Once any of

these standard errors is calculated, its application to testing for mediation is identical. A 95% confidence interval for the mediated effect is calculated as $b_1b_2b_3$ plus and minus 1.96 times the standard error (where 1.96 is the critical value from the normal distribution). If the confidence interval does not include zero, the null hypothesis is rejected.

Difference-in-coefficients tests. In the single-mediator case, an estimator of the mediated effect may be obtained by taking the difference between the coefficient relating the independent and dependent variables before and after adjustment for the mediator. There is not a clear analogous difference-in-coefficient test for the three-path mediated effect, illustrating the problem with extending the difference in coefficient tests to more complicated models. For example, the difference between the coefficient relating the predictor to the second mediator could be examined before and after adjustment for the first mediator, and the coefficient relating the first mediator to the outcome could be examined before and after adjustment for the second mediator. Although it may be possible to develop a three-path test of mediation based on differences in coefficients, this method would likely be cumbersome in comparison to the product-of-coefficients test. As a result, difference-in-coefficients tests for mediation are not considered in the present study.

Resampling methods. In discussing the single-mediator model, MacKinnon et al. (2002) noted that a major difficulty in using product-of-coefficients tests for mediation is that the distribution of the product of two regression coefficients is not normal as the tests assume. In the three-path situation, $b_1b_2b_3$ estimates the mediated effect, and its distribution is also nonnormal (Craig, 1936; Springer & Thompson, 1970), leading to poor performance of the product-of-coefficients tests. In situations such as this, where the assumptions of classical statistics are violated, resampling methods often perform better because they do not make as many problematic assumptions (Manly, 1997). Bootstrapping is one such resampling method that has been widely applied to cases in which classical methods do not perform well. Bootstrapping involves drawing a large number of samples with replacement from the original sample. Sampling with replacement means that the bootstrap samples, although all the same size as the original sample, can exclude some cases from the original sample and include duplicates of others. The model of interest is estimated in each bootstrap sample as in the original data. The distribution of sample statistics estimated in each bootstrap sample can be used to perform significance tests or to form confidence intervals.

MacKinnon et al. (2004) used a Monte Carlo study to compare the performance of several resampling methods, including several variants of the bootstrap, in the single-mediator case. Among the best performers they found were the percentile bootstrap and the bias-corrected bootstrap. Bollen and Stine (1990) used the percentile and bias-corrected bootstrap methods to estimate confidence intervals for real data in the single-mediator case. They found that particularly in their small sample, the bootstrap captured the asymmetry in the sampling distribution missed by the product-of-coefficients test using the multivariate delta standard error. Shrout and Bolger (2002) also demonstrated and advocated for the use of the bootstrap in testing mediational models.

Bootstrap methods can be generalized from the two-path to the three-path situation. In the three-path situation, regression models are first estimated for the original data to find the coefficients b_1, b_2, and b_3. A large number of bootstrap samples are drawn, the same

models are estimated for each bootstrap sample, and the $b_1b_2b_3$ estimates from each bootstrap sample are used to form the bootstrap distribution. The limits of a percentile bootstrap confidence interval are simply the values of $b_1b_2b_3$ at the $\alpha/2$ and $1 - \alpha/2$ percentiles of the bootstrap distribution, where α is the nominal Type I error rate. For example, for the typical $\alpha = .05$, the limits are the 2.5th and 97.5th percentiles of the distribution. The bias-corrected confidence interval limits are also taken from the bootstrap distribution, but they are adjusted if the bootstrap distribution fails to center at the sample estimate of the mediated effect. Details of the bias correction procedure are given in Table 1, in Efron and Tibshirani (1993, chap. 14) and in Manly (1997, sec. 3.4).

An Empirical Example

Six methods of testing the three-path mediated effect have been described. These methods include one causal-steps method, the joint significance test; three product-of-coefficients methods, the multivariate delta variance estimator, the unbiased variance estimator, and the exact variance estimator; and two bootstrap methods, the percentile bootstrap and the bias-corrected bootstrap. These six methods were applied to data based on Kiefer's (2005) study of the role of negative emotion in the effects of organizational change. Two of her hypotheses may be considered together as a three-path mediational model. First (Hypothesis 1a), the effect of organizational change on employees' negative emotions was expected to be mediated by working conditions. Second (Hypothesis 2b), negative emotions were expected to predict withdrawal from the organization. In terms of Figure 1 and Equations 1-3, organizational change is X, working conditions is M_1, negative emotions is M_2, and withdrawal is Y. The paths from X to M_1 (b_1) and from M_1 to M_2 (b_2) were expected to be negative, and the path from M_2 to Y (b_3) was expected to be positive.

The application of these methods was not expected to produce exactly the same results as Kiefer's (2005) for three reasons. First, her entire model was more complex, including three other variables that were not included in the present analysis. Second, as the bootstrap methods require case-level data and we did not have Kiefer's original data, our analyses are based on a simulated data set of the same size as the original data ($N = 155$), simulated to match as closely as possible the variances and correlations of Kiefer's (Table 1) data. Third, Kiefer reported standardized results, and the present application reports unstandardized results for simplicity.

Equations 1-3 were estimated using ordinary least squares (OLS) regression for the data based on Kiefer (2005). As shown in Table 2, each of b_1, b_2, and b_3, the coefficients making up the mediated effect, was significantly nonzero. All three effects were in the predicted direction: Organizational change was negatively related to working conditions, working conditions were negatively related to negative emotion, and negative emotion was positively related to withdrawal. The significance of all three coefficients means that the joint significance test rejected the null hypothesis of no mediation.

The estimate of the mediated effect for the product of coefficients methods, $b_1 \times b_2 \times b_3$, was 0.332. Standard errors of the mediated effect based on the three different estimators are shown in Table 2. The estimators produced very similar results because the coefficients were large and their standard errors small. Under these circumstances, the formulas for all three estimators (see Table 1) are dominated by the first three terms,

Table 2
Results for Each Method Applied to Data Based on Kiefer (2005)

Method	Effect	*SE*	Test	Hypothesis_0 Test Result
Joint significance	$b_1 = -0.947$	$s_{b1} = 0.299$	$t(153) = -3.17$, $p = .002$	
	$b_2 = -0.639$	$s_{b2} = 0.062$	$t(152) = -10.35$, $p < .001$	
	$b_3 = 7.06$	$s_{b3} = 0.078$	$t(151) = 7.06$, $p < .001$	Reject Hypothesis_0
Multivariate delta standard error	$b_1b_2b_3 = 0.332$	$s_{\text{multivariate delta}} = 0.119$	95% CI = [0.098, 0.565]	Reject Hypothesis_0
Unbiased standard error	$b_1b_2b_3 = 0.332$	$s_{\text{unbiased}} = 0.118$	95% CI = [0.101, 0.563]	Reject Hypothesis_0
Exact standard error	$b_1b_2b_3 = 0.332$	$s_{\text{exact}} = 0.121$	95% CI = [0.095, 0.568]	Reject Hypothesis_0
Percentile bootstrap	$b_1b_2b_3 = 0.332$	—	95% CI = [0.109, 0.626]	Reject Hypothesis_0
Bias-corrected bootstrap	$b_1b_2b_3 = 0.332$	—	95% CI = [0.131, 0.678]	Reject Hypothesis_0

Note: The standard error of $b_1b_2b_3$ was not estimated using the bootstrap methods. CI = confidence interval.

which they all have in common, whereas the remaining terms, which are present only in the unbiased and exact standard errors, are very small. Confidence intervals calculated using each estimator are also shown in Table 2. These were calculated by multiplying each estimated standard error by 1.96 (the critical z value for a two-tailed test with $\alpha = .05$) and adding and subtracting the result from the estimate of the mediated effect. None of the confidence intervals included zero, so the null hypothesis of mediation was rejected for all three variance estimators.

Results for the bootstrap methods are shown in Table 2. For the percentile bootstrap, 1,000 bootstrap samples were drawn from the original sample. For each bootstrap sample, the regressions in Equations 1-3 were estimated, and the estimate of the mediated effect $b_1b_2b_3$ was also calculated. The limits of the 95% confidence interval for the percentile bootstrap are the 2.5th and 97.5th percentiles of the bootstrap distribution. Therefore, the limits of the percentile bootstrap confidence interval in Table 2 were found by simply ordering the 1,000 bootstrap estimates of $b_1b_2b_3$ and picking the 25th and the 975th values. As the interval did not include zero, the null hypothesis of no mediation was rejected.

The bias-corrected bootstrap adjusts which percentiles are chosen from the bootstrap distribution based on whether the bootstrap distribution is mostly above or below the original sample estimate of $b_1b_2b_3$. The steps are listed in Table 1. The proportion of bootstrap estimates of $b_1b_2b_3$ greater than the original sample estimate was .442. The z-score corresponding to this probability, that is, the z-score above which .442 of a standard normal distribution falls, is 0.1459. Two times this z-score, 0.2918, was added and subtracted from the usual two-tailed 95% z critical value of ± 1.960 to yield $z_{\text{lower}} = -1.668$ and $z_{\text{upper}} = 2.252$. These z scores represent the 4.8th and the 98.8th percentiles of a standard

normal distribution. The adjustment of the bias-corrected bootstrap can be seen clearly in this example. The majority of the bootstrap estimates of $b_1b_2b_3$ fell below the original sample estimate—442 fell above and 558 fell below—so the percentiles for finding the limits of the interval were shifted up: from the 2.5th to the 4.8th percentile at the low end and from the 97.5th to the 98.8th percentile at the high end. Therefore, the limits of the bias-corrected bootstrap confidence interval are the 48th and the 988th estimates from the ordered distribution of 1,000 bootstrap estimates of $b_1b_2b_3$. As for the percentile bootstrap, the confidence interval did not include zero, so the null hypothesis of no mediation was rejected.

In summary, the six methods all suggested the same conclusion: Reject the null hypothesis of no mediation. The different methods did not produce exactly the same results, though. The standard error formulas for the product-of-coefficients methods and the bootstrapping methods all yielded different confidence intervals. These differences raise the question of whether the different methods will yield similar results across a broader range of circumstances.

To answer this question, a Monte Carlo study was performed. Its purpose was to compare the performance of these methods in different sample sizes and for different sizes of the coefficients making up the mediated effect. The methods were compared in terms of (a) relative bias of their estimates of the standard error of the mediated effect (for the product-of-coefficients methods only), (b) their Type I error, (c) their power, and (d) the coverage of their confidence intervals (for all methods except the joint significance test).

Method

Both simulation and analysis of the data were done using Statistical Analysis System (SAS) software (SAS Institute, 2005). Data were simulated using Equations 1, 2, and 3. The X variable was simulated to either be normally distributed with a mean of 0 and a variance of 1 or to be dichotomous with a .5 probability of being in each category. The latter condition was intended to mimic an experimental design with two groups. The residuals ε_1, ε_2, and ε_3 were simulated to be normally distributed with a mean of zero and a variance of 1. The residuals and X when continuous were simulated using the SAS RANNOR function. When dichotomous, X was simulated using RANNOR and then dichotomized at 0 into 0 and 1. The intercepts β_{01}, β_{02}, and β_{03} were set to zero. The sizes of the coefficients β_1, β_2, β_3, and β_4 were, in different conditions of the simulation, set equal to zero or set to correspond to Cohen's (1988) small (.14), medium (.39), or large (.59) partial correlations. Because of the large computational demands of the bootstrap methods, the number of conditions was reduced by including only those in which β_2 and β_3 are no larger than β_1, meaning that 30 of a possible 64 conditions defined by β_1, β_2, β_3 were included. These conditions were chosen because, particularly in experimental studies where X is manipulated to create maximal effect on M_1, it seems reasonable to expect that β_1 will be the largest path in the mediational chain. All levels of β_4, the direct effect, were included in the simulation, meaning that both full mediation ($\beta_4 = 0$) and partial mediation ($\beta_4 \neq 0$) models were studied. The relations represented by β_5 and β_6 were expected to be independent of the β_1, β_2, β_3 paths that define the mediation relation. To verify this result, β_5 and β_6 were set to zero or large (.59), but the corresponding coefficients b_5 and b_6 were estimated in all conditions, regardless of their corresponding true values. Sample size was set to 50, 100, 200,

500, or 1,000 in different conditions. There were 4,800 total conditions: two levels of distribution of $X \times 30$ levels of β_1, β_2, and β_3 in which $\beta_2 \leq \beta_1$ and $\beta_3 \leq \beta_1 \times 4$ levels of $\beta_4 \times 2$ levels of $\beta_5 \times 2$ levels of $\beta_6 \times 5$ levels of sample size. One thousand replications were run for each condition. For each replication in each condition, all six methods were used to test for mediation. For the bootstrap methods, 1,000 bootstrap samples were drawn in each replication.

Performance of Standard Error Estimators

The product-of-coefficients methods were compared using the relative bias of their standard errors. Relative bias of each estimated standard error was calculated for each replication in a condition as

$$\text{Relative Bias} = \frac{SE_{b1b2b3} - SD_{b1b2b3}}{SD_{b1b2b3}}, \tag{5}$$

where SE_{b1b2b3} is the estimated standard error of $b_1b_2b_3$ calculated using one of the three estimators for one replication within a condition, and SD_{b1b2b3} is the standard deviation of $b_1b_2b_3$ in the 1,000 replications for the same condition. The latter quantity estimates the true standard error of the mediated effect for the condition. The numerator of Equation 4 is the bias of a standard error estimate. Dividing by the estimated true standard error puts relative bias in the metric of bias as a proportion of the true value, which allows for easier comparison across conditions. The relative bias for each of the three methods in the condition as a whole was calculated as the mean of the individual replications' relative bias values. Based on Kaplan's (1988) recommendation, a mean relative bias of .10 or smaller in absolute value was considered small enough to make the standard error estimator usable in an applied context.

Type I Error and Power

The six methods of testing for mediation in the three-path situation were each used to test the null hypothesis of no mediation for each of the replications in each condition. The nominal Type I error rate was set to .05 for all methods. The proportion of replications for which the null hypothesis was rejected was calculated. This proportion was the Type I error rate in conditions in which the null hypothesis was true (i.e., $\beta_1\beta_2\beta_3 = 0$). It was the empirical power level in conditions in which the null hypothesis was false (i.e., $\beta_1\beta_2\beta_3 \neq 0$). For the joint significance test, testing the null hypothesis of no mediation required a separate hypothesis test for each of b_1, b_2, and b_3. If all three null hypotheses were rejected, the null hypothesis of no mediation was rejected. For all other methods, a 95% confidence interval around the mediated effect was estimated. If this interval did not include zero, the null hypothesis of no mediation was rejected (see Table 1).

Coverage

The five methods that were used to generate confidence intervals were compared in terms of their coverage. (The joint significance test was not used to estimate confidence

intervals.) Coverage was the proportion of replications in which the confidence interval included the true mediated effect, $\beta_1\beta_2\beta_3$. As 95% confidence intervals were estimated, the methods performed well when their coverage levels were near .95.

Results

The same modeling approach was used to analyze results for each of the four study outcomes: relative bias of standard errors, Type I error, power, and coverage. As described above, the conditions in the study were defined by eight factors: distribution of X, size of β_1-β_6, and sample size. The six methods of testing for mediation constituted a within-participant factor, as each of the methods was applied to all of the replications in each condition. In models of the study outcomes, method was treated as a between-participants factor, making for a total of nine study factors. This treatment of the method factor did not affect the model results, as the between-participants versus within-participant distinction is important only in calculating error terms and significance tests, and as described below, model results were interpreted using effect sizes rather than significance tests. A separate model was estimated for each of the study outcomes. The intent was to model each outcome using all nine factors and their interactions. Because of the size of the data sets (as many as 24,000,000 cases for studying coverage, for example) and the size of the design matrix with all nine study factors and all possible interactions among them, the models could not be estimated when all factors were treated as categorical. To be able to test these interactions, the models were estimated with the quantitative predictors—sample size and β_1 through β_6—entered as numerical values rather than as categorical predictors. These predictors were centered before interaction terms were constructed, following the recommendation of Aiken and West (1991). Entering these factors as numerical values traded the multiple-degree-of-freedom test for differences of any form across levels of the factors for a single-degree-of-freedom test for the linear trend in each factor. As effects of these factors were expected to be at least monotonic, even if not precisely linear, this seemed to be a reasonable way of retaining the higher order interactions in the model. Finally, the type of model depended on the form of the outcome being modeled. As relative bias is continuous, it was modeled using an analysis of covariance (ANCOVA). The other outcomes, Type I error, power, and coverage, are dichotomous (i.e., reject versus fail to reject the null hypothesis, confidence interval does versus does not include the true value), so they were modeled using logistic regression, although the predictors were entered just as they were for the ANCOVA model of relative bias.

As the number of cases for these models was so large, power to detect even trivially small effects was very high. Therefore, rather than interpret effects based on the conventional $p < .05$ significance criterion, or any other significance criterion, effects were interpreted on the basis of their associated effect sizes. In the ANCOVA model of relative bias, the effect size measure was the proportion of variance accounted for, ω^2. In logistic regression, there is no true proportion of variance accounted for measure, but there are analogs. Following the recommendation of Menard (2000), R_L^2 was used for the logistic models of Type I error, power, and coverage. R_L^2 tells the proportion of reduction in deviance (badness of fit) in a model attributable to a predictor. For ω^2, an effect accounting for at least 1% of

the variance was considered large enough to be interpretable. As effect size measures in logistic regression tend to be smaller than R^2 or ω^2 values from OLS regression or analysis of variance (Hagle & Mitchell, 1992; Hosmer & Lemeshow, 2000, p. 167), the threshold for interpretation for R_L^2 was set to 0.5%. Effects meeting or exceeding the 1% threshold for ω^2 or the 0.5% threshold for R_L^2 are referred to below as being "nontrivial." To reduce the size of the model summary tables, only nontrivial effects are listed separately. Tables of cell means for each study are constructed to highlight nontrivial effects and generally collapse across effects not reaching this threshold.

One other issue that was important in modeling all of the outcomes was that the unbiased variance estimator sometimes produced negative estimates of the variance of $b_1b_2b_3$. MacKinnon et al. (2002) had similar results for the unbiased estimator in the single-mediator case. As can be seen from the formula in Table 1, this result occurs because, unlike the multivariate delta and exact estimators, the unbiased estimator subtracts some terms to reduce positive bias of the estimator. In conditions in which the values of β_1, β_2, and β_3 were small or zero, sampling error often made the sum of these subtracted terms larger than the sum of the added terms. For example, in conditions in which two or three of β_1, β_2, and β_3 were zero, 30% to 40% of the unbiased variance estimates were negative across sample sizes. In conditions in which two or three of β_1, β_2, and β_3 were nonzero, on the other hand, negative estimates occurred in as many as 30% of replications, but only for the smallest effect sizes and sample sizes. For larger effect sizes and sample sizes, the percentage of negative variance estimates quickly dropped to near zero. The tendency of the unbiased variance estimator to sometimes yield negative values even though its true value cannot be negative makes it similar to the adjusted R^2 in OLS regression. The adjusted R^2 also has a term subtracted to avoid positive bias, and under conditions where its true value is near or equal to zero, sampling error often makes it slightly negative. Negative unbiased variances are problematic because when the square root is taken, the estimated standard error is an imaginary number, which cannot be used to construct a confidence interval. In the analyses of study outcomes, unbiased variance estimate results were included only if they were nonnegative.

Relative Bias of Standard Error Estimators

The multivariate delta, unbiased, and exact estimators of the variance of $b_1b_2b_3$ were compared in terms of their relative bias. As shown in Table 3, study factors collectively accounted for 14.6% of the variance in relative bias. Individual effects are not listed because no individual effect accounted for as much as 1% of the variance. Although no individual effects accounted for nontrivial variance, the set of terms including method, taken together, accounted for nontrivial variance ($\omega^2 = .016$), as did the sets of terms including β_2 ($\omega^2 = .024$) and β_3 ($\omega^2 = .023$; note that these sets of terms are overlapping). Table 4 therefore presents mean relative bias values as a function of method and size of the mediated effect, collapsing across levels of β_4, β_5, β_6, and distribution of X. To save space, mediated effect sizes are also collapsed across conditions in which results were similar. The methods' performance depended on the number of zero paths in null true models and on the size of the mediated effect in null false models, so the rows of the table are defined by these groups. The table also displays results separately by sample size, as there

Table 3
Relative Bias Model Summary

Effect	SS	*df*	*MS*	*F*	*p*	ω^2
Model	579395	767	755.4	3,104.6	<.001	.146
Error	3398638	13967905	0.2			
Total	3978033	13968672				

Note: Effects are not listed separately because no individual effect accounted for as much as 1% of the variance. Results for the unbiased standard error are based on only replications in which its variance estimate was nonnegative.

appeared to be a consistent, if small, effect of this factor. For each table entry, which is a mean relative bias collapsing across a number of conditions, the proportion of the conditions in which the absolute value of the mean relative bias exceeded .10 is also shown in parentheses.

Results are discussed in terms of comparing methods, as this is our major focus, but it should be kept in mind that as no effects accounted for as much as 1% of the variance, differences were generally small. The exact variance estimator performed most poorly of the methods, almost always being biased high. Its relative bias exceeded 1 in conditions where $\beta_1 = \beta_2 = \beta_3 = 0$ and only declined to below .10 on average when $\beta_1\beta_2\beta_3$ exceeded .01. In conditions in which two or three of β_1, β_2, and β_3 were zero, its relative bias was unaffected by sample size, but in conditions in which two or three of β_1, β_2, and β_3 were nonzero, its relative bias improved with increasing sample size. The multivariate delta variance estimator had a similar pattern of performance, but with much less relative bias in all conditions. For example, its worst relative bias was only .30, and its mean relative bias was below .10 in nearly all conditions in which two or three of β_1, β_2, and β_3 were nonzero. The unbiased variance estimator differed from the other two in that it was more often negatively than positively biased. It performed quite well when $\beta_1 = \beta_2 = \beta_3 = 0$; was positively biased when two of β_1, β_2, and β_3 were zero; and was negatively biased when two or three of β_1, β_2, and β_3 were nonzero. The reality is that the unbiased variance estimator is always negatively biased; as described above, these results are based on only the replications in which the unbiased variance estimate was nonnegative so its square root could be taken to get a standard error. If all replications in which the unbiased variance was negative had been entered as a zero standard error estimate, for example, it would likely have exhibited strong negative bias in the conditions where two or three of β_1, β_2, and β_3 were zero.

Type I Error

All six methods of testing for mediation were compared in terms of their Type I error. The $\beta_2 \times \beta_3$ interaction and all higher order interactions including both coefficients were excluded from the model, as $\beta_2 \times \beta_3 = 0$ in all null hypothesis true conditions. As shown in Table 5, study factors accounted for a 10.4% deviance reduction. The method of testing for mediation had by far the largest effect ($R_L^2 = .030$), and sample size also had a

Table 4
Mean Relative Bias of Standard Errors and Proportions of Conditions in Which Absolute Mean Relative Bias |RB| > .10

	Sample Size									
	$n = 50$		$n = 100$		$n = 200$		$n = 500$		$n = 1{,}000$	
$\beta_1, \beta_2, \beta_3$	Rel. Bias	($\|RB\| > .10$)	Rel. Bias	($\|RB\| > .10$)	Rel. Bias	($\|RB\| > .10$)	Rel. Bias	($\|RB\| > .10$)	Rel. Bias	($\|RB\| > .10$)
					Mutivariate delta standard error					
All paths = 0	0.230	(0.906)	0.266	(0.906)	0.284	(1.000)	0.308	(0.938)	0.275	(0.938)
Two paths = 0	0.192	(0.885)	0.222	(0.938)	0.223	(0.979)	0.224	(0.990)	0.243	(1.000)
One path = 0	0.041	(0.188)	0.023	(0.063)	0.009	(0.021)	0.004	(0.000)	0.001	(0.000)
$\beta_1, \beta_2, \beta_3 < .01$	0.030	(0.078)	0.000	(0.000)	−0.014	(0.000)	−0.004	(0.000)	−0.003	(0.000)
$\beta_1, \beta_2, \beta_3 \geq .01$	−0.007	(0.005)	−0.007	(0.000)	−0.004	(0.003)	−0.001	(0.000)	−0.001	(0.000)
					Unbiased standard error					
All paths = 0	−0.025	(0.188)	−0.002	(0.156)	0.010	(0.125)	0.031	(0.219)	0.004	(0.125)
Two paths = 0	0.042	(0.354)	0.103	(0.479)	0.144	(0.667)	0.175	(0.822)	0.216	(0.865)
One path = 0	−0.061	(0.247)	−0.056	(0.122)	−0.052	(0.102)	−0.038	(0.044)	−0.024	(0.023)
$\beta_1, \beta_2, \beta_3 < .01$	−0.095	(0.406)	−0.104	(0.563)	−0.102	(0.516)	−0.054	(0.047)	−0.028	(0.000)
$\beta_1, \beta_2, \beta_3 \geq .01$	−0.077	(0.266)	−0.053	(0.073)	−0.030	(0.005)	−0.012	(0.000)	−0.006	(0.000)
					Exact standard error					
All paths = 0	1.248	(1.000)	1.328	(1.000)	1.352	(1.000)	1.414	(1.000)	1.346	(1.000)
Two paths = 0	0.826	(1.000)	0.807	(1.000)	0.740	(1.000)	0.681	(1.000)	0.679	(1.000)
One path = 0	0.334	(0.826)	0.217	(0.641)	0.127	(0.484)	0.057	(0.167)	0.030	(0.065)
$\beta_1, \beta_2, \beta_3 < .01$	0.418	(1.000)	0.236	(0.953)	0.112	(0.516)	0.047	(0.047)	0.022	(0.000)
$\beta_1, \beta_2, \beta_3 \geq .01$	0.098	(0.369)	0.047	(0.133)	0.024	(0.023)	0.010	(0.000)	0.005	(0.000)

Note: For each sample size, entries in the left column are mean relative bias values, collapsing across study factors for which results were similar. Entries in the right column (in parentheses) are the proportions of the collapsed conditions in which the absolute value of the mean relative bias was greater than .10, the cutoff suggested by Kaplan (1988). Study factors collapsed because of similar results are *X* distribution, β_4, β_5, and β_6. The 30 levels of β_1, β_2, and β_3 are also collapsed into the 5 levels listed for each method. Results for the unbiased standard error are based on only replications in which its variance estimate was nonnegative. Rel. = relative.

nontrivial effect ($R^2_L = .006$). Table 6 therefore shows Type I error rates for the different methods (although the product-of-coefficients methods are averaged because they performed very similarly) by sample size. The table is also broken down by size of β_1, β_2, and β_3, as effects of β_2 and β_3 approached the threshold for being nontrivial ($R^2_L = .004$ and .003), and inspection of cell means suggested that Type I error rates varied meaningfully with the sizes of these coefficients. The 30 conditions defined by β_1, β_2, and β_3 are collapsed into 4 based on the size of the product of the two largest coefficients (the null hypothesis is true when any of β_1, β_2, and β_3 is zero, so this product was nonzero in some conditions). As excess Type I error was of particular concern, for each entry in the table, the proportion of the conditions on which it is based in which a 95% confidence interval for the Type I error rate was completely above .05 is also shown in parentheses.

Type I error was lower than the nominal level across methods in conditions where at least two of β_1, β_2, and β_3 were zero (in the first row for each condition). In conditions where only one of β_1, β_2, and β_3 was zero, Type I error increased with increasing size of the nonzero coefficients. For the product-of-coefficients methods, the mean Type I error rate never reached the nominal level, even for the largest effect sizes and sample sizes. Their Type I error rate did not significantly exceed .05 in any condition. Type I error rates for the joint significance test and percentile bootstrap were higher than for the product-of-coefficients methods. For both, the Type I error rate did approach the nominal level for $n = 500$ or 1,000 in conditions where the product of the two nonzero coefficients exceeded .05 and for $n = 100$ or more in conditions where the product of the nonzero coefficients exceeded .10. Neither method had Type I error significantly exceeding its nominal level in more than 4.7% of cells, though. The bias-corrected bootstrap's Type I error rate fell below the nominal rate in conditions in which two or three of β_1, β_2, and β_3 were zero, but increased to beyond the nominal rate as the size of the nonzero effects and sample size increased. For smaller products of nonzero coefficients, the bias-corrected bootstrap's Type I error rate got worse with increasing sample size, reaching a maximum of .081, with 81.3% of cells significantly exceeding .05. For larger products of nonzero coefficients, its Type I error rate reached a maximum of .087, with 89.1% of cells significantly exceeding .05, but improved with increasing sample size.

Power

All six methods of testing for mediation were compared in terms of their power. As shown in Table 7, study factors accounted for a 58.4% deviance reduction. The largest effect was that of sample size ($R^2_L = .244$); smaller nontrivial amounts of deviance reduction were associated with β_2, β_3, their interaction, their interactions with sample size, and the three-way interaction. As with models of relative bias and Type I error, there were no effects of β_4, β_5, or β_6. Unlike the models of other study outcomes, though, no nontrivial effects included method. Table 8 shows power levels as a function of size of the mediated effect, sample size, distribution of X, and method (the product-of-coefficients methods are averaged because they performed very similarly). Distribution of X is included as a factor in the table because it approached the threshold for being nontrivial ($R^2_L = .004$), and inspection of cell means suggested that power differences across this factor were large enough to be important. Method is also included as a factor in the table in spite of not

Table 5
Type I Error Model Summary, Nontrivial Effects ($R_L^2 \geq .005$)

Effect	LR χ^2	*df*	*p*	R_L^2
Method	97,620.3	5	<.001	.030
Sample size	18,661.6	1	<.001	.006
All other effects ($R_L^2 < .005$)	204,813.9	1145	<.001	.063
Model	338,898.8	1151	<.001	.104

Note: The sum of individual effects' R_L^2 differs from the overall model R_L^2 because R_L^2 is not a true proportion of variance accounted for measure. Results for the unbiased standard error are based on only replications in which its variance estimate was nonnegative. LR = likelihood ratio.

appearing in any nontrivial effects because the primary purpose of this study is to compare methods of testing for mediation. The set of all terms including method did, collectively, reduce deviance at a nontrivial level (likelihood ratio $\chi^2[1, 280] = 556, 707.7$, $p < .001$, $R_L^2 = .031$). By comparison, the set of terms including β_4, β_5, or β_6, although conventionally significant, did not approach the criterion for being nontrivial (likelihood ratio $\chi^2[1, 344] = 1, 762.3$, $p < .001$, $R_L^2 = .0001$; note that the two sets of terms are overlapping).

The main effects of sample size, size of the mediated effect, and distribution of X were straightforward: Power increased with increasing sample size, with increasing size of the mediated effect, and was greater for a continuous normal X than for a dichotomous X. Sample size moderated differences across mediated effect size. In smaller samples, increasing effect size was always associated with increasing power. In larger sample size conditions, though, a ceiling effect occurred where power approached 1 for smaller and smaller effects, making the size of the mediated effect much less a predictor of power in the largest sample sizes than in smaller sample sizes.

The effect of method also varied with sample size and effect size. For the largest sample sizes and effect sizes, power levels for all methods approached 1, and differences between methods were near zero. For smaller sample sizes and effect sizes, though, differences between methods were larger. The bias-corrected bootstrap had the greatest power in all conditions in which power differences were manifest. It was followed by the joint significance test and the percentile bootstrap, which had similar power levels. The product-of-coefficients methods had the least power.

Coverage

The five methods used to generate confidence intervals for the mediated effect (the joint significance test cannot be easily used to estimate confidence intervals) were compared in terms of their coverage. As shown in Table 9, study factors accounted for a 4.7% deviance reduction. Only method, β_2, and β_3 reduced a nontrivial amount of deviance in the model.

Coverage in null hypothesis true conditions can be inferred from Type I error. This is because when the null hypothesis is true, the true value that a confidence interval should include is zero. If zero is not included, the null hypothesis is rejected, and the interval has failed to capture the true value. If zero is included, the null hypothesis is retained, and the

Table 6
Type I Error Rates and Proportion of Conditions in Which Type I Error Significantly Exceeded .05

	Sample Size									
	$n = 50$		$n = 100$		$n = 200$		$n = 500$		$n = 1{,}000$	
Maximum of $\beta_1\beta_2, \beta_1\beta_3, \beta_2\beta_3$	Type I Error	$p > .05$	Type I Error	$p > .05$	Type I Error	$p > .05$	Type I Error	$p > .05$	Type I Error	$p > .05$
					Joint significance test					
max = 0	.001	(.000)	.001	(.000)	.002	(.000)	.002	(.000)	.002	(.000)
0 < max < .05	.001	(.000)	.003	(.000)	.008	(.000)	.026	(.000)	.039	(.000)
.05 ≤ max < .10	.005	(.000)	.011	(.000)	.023	(.000)	.043	(.000)	.050	(.008)
.10 ≤ max	.028	(.000)	.041	(.000)	.048	(.000)	.050	(.005)	.051	(.021)
					Product-of-coefficients methods (averaged)					
max = 0	.006	(.000)	.006	(.000)	.006	(.000)	.006	(.000)	.006	(.000)
0 < max < .05	.007	(.000)	.005	(.000)	.004	(.000)	.002	(.000)	.007	(.000)
.05 ≤ max < .10	.005	(.000)	.003	(.000)	.003	(.000)	.009	(.000)	.023	(.000)
.10 ≤ max	.004	(.000)	.009	(.000)	.022	(.000)	.037	(.000)	.044	(.000)
					Percentile bootstrap					
max = 0	.001	(.000)	.001	(.000)	.001	(.000)	.001	(.000)	.001	(.000)
0 < max < .05	.000	(.000)	.001	(.000)	.005	(.000)	.020	(.000)	.036	(.000)
.05 ≤ max < .10	.004	(.000)	.008	(.000)	.018	(.000)	.040	(.000)	.051	(.000)
.10 ≤ max	.027	(.005)	.043	(.026)	.051	(.047)	.052	(.031)	.052	(.031)
					Bias-corrected bootstrap					
max = 0	.005	(.000)	.005	(.000)	.005	(.000)	.005	(.000)	.005	(.000)
0 < max < .05	.006	(.000)	.012	(.000)	.028	(.000)	.066	(.500)	.081	(.813)
.05 ≤ max < .10	.020	(.000)	.034	(.000)	.055	(.102)	.081	(.922)	.078	(.922)
.10 ≤ max	.074	(.672)	.087	(.891)	.082	(.932)	.066	(.469)	.060	(.250)

Note: For each sample size, entries in the left column are mean Type I error rates, collapsing across study factors for which results were similar. Entries in the right column (in parentheses) are the proportions of the collapsed conditions in which the Type I error rate significantly exceeded .05. Each Type I error rate was tested by forming a 95% confidence interval. As each condition consisted of 1,000 replications, a Type I error rate of .066 (which has a 95% confidence interval of [.051, .081]) or greater significantly exceeded .05. Study factors collapsed because of similar results are X distribution, β_4, β_5, and β_6. The 16 levels of β_1, β_2, and β_3 in which the null hypothesis is true are also collapsed into the 4 levels listed for each method. Results for the product-of-coefficients methods are collapsed, including for the unbiased standard error–only replications in which its variance estimate was nonnegative.

Table 7
Power Model Summary, Nontrivial Effects ($R^2_L \geq .005$)

Effect	LR χ^2	*df*	*p*	R^2_L
Sample size	4,460,920.4	1	$\leq .001$	.244
β_2	847,728.3	1	$\leq .001$	.046
β_3	841,959.7	1	$\leq .001$	.046
Sample Size $\times$ β_2	338,318.9	1	$\leq .001$	.019
Sample Size $\times$ β_3	335,062.0	1	$\leq .001$	.018
$\beta_2 \times \beta_3$	445,095.5	1	$\leq .001$	.024
Sample Size $\times$ $\beta_2 \times \beta_3$	276,536.8	1	$\leq .001$	.015
All other effects ($R^2_L \leq .005$)	2,983,032.0	1528	$\leq .001$	.163
Model	10,652,891.0	1535	$\leq .001$	.584

Note: The sum of individual effects' R^2_L differs from the overall model R^2_L because R^2_L is not a true proportion of variance accounted for measure. Results for the unbiased standard error are based on only replications in which its variance estimate was nonnegative. LR = likelihood ratio.

interval has succeeded in capturing the true value. Therefore, coverage is equal to 1 minus Type I error in null true conditions. Returning to Table 6, it is clear that all methods had too high coverage (indicating too wide confidence intervals) in conditions in which two or three of β_1, β_2, and β_3 were zero. Coverage fell to near the nominal level in the largest effect size and sample size conditions for the product-of-coefficients methods. For the percentile bootstrap, coverage fell to near the nominal level for somewhat smaller effect and sample sizes. The bias-corrected bootstrap had coverage that fell as low as .913 for the largest effect sizes in $n = 100$ conditions and as low as .919 for smaller effect sizes and larger sample sizes.

Table 10 shows coverage in null false conditions and, similar to the Type I error results, the proportion of cells in which coverage was significantly below .95 as a function of method, size of the mediated effect, and sample size. Sample size is included, although its effect did not reach the threshold for being nontrivial, because there nevertheless appeared to be a clear pattern of results across its levels. Unlike Tables 6 and 8, Table 10 presents the product-of-coefficients methods separately because their coverage levels were quite different. The unbiased standard error had the lowest coverage levels (indicating too narrow confidence intervals), with a minimum of .839, with 100% of conditions falling significantly below .95. Its coverage levels increased with increasing sample size and effect size. The multivariate delta standard error had a similar pattern of performance but with generally better coverage; its lowest coverage was .880. The exact standard error had the best coverage of the product-of-coefficients methods. It was as low as .920 in one condition but fell between .940 and .950 in most conditions where the mediated effect was .01 or larger and sample size was at least 200. The percentile bootstrap had very good coverage performance, with a minimum of .930, and at least .940 across conditions when sample size was at least 100. The bias-corrected bootstrap had too low coverage, with a minimum of .897 in the smallest sample size and effect size conditions, but it matched the percentile bootstrap in having good coverage for sample sizes of at least 200.

Table 8
Power Levels

	Sample Size and Distribution of X									
	$n = 50$		$n = 100$		$n = 200$		$n = 500$		$n = 1{,}000$	
$\beta_1\beta_2\beta_3$	$X \sim 0\ 1$	$X \sim$ normal	$X \sim 0\ 1$	$X \sim$ normal	$X \sim 0\ 1$	$X \sim$ normal	$X \sim 0\ 1$	$X \sim$ normal	$X \sim 0\ 1$	$X \sim$ normal
					Joint significance test					
$\beta_1\beta_2\beta_3 < .01$	0.004	0.010	0.021	0.046	0.114	0.190	0.514	0.721	0.792	0.980
$.01 \le \beta_1\beta_2\beta_3 < .05$	0.048	0.101	0.174	0.240	0.427	0.463	0.856	0.858	0.992	0.992
$.05 \le \beta_1\beta_2\beta_3 < .10$	0.207	0.465	0.609	0.910	0.883	0.999	0.995	1.000	1.000	1.000
$.10 \le \beta_1\beta_2\beta_3$	0.409	0.764	0.807	0.974	0.984	1.000	1.000	1.000	1.000	1.000
					Product-of-coefficients methods (averaged)					
$\beta_1\beta_2\beta_3 < .01$	0.006	0.005	0.004	0.005	0.007	0.029	0.199	0.373	0.640	0.926
$.01 \le \beta_1\beta_2\beta_3 < .05$	0.004	0.019	0.033	0.110	0.216	0.359	0.768	0.816	0.987	0.989
$.05 \le \beta_1\beta_2\beta_3 < .10$	0.020	0.118	0.272	0.706	0.807	0.997	0.994	1.000	1.000	1.000
$.10 \le \beta_1\beta_2\beta_3$	0.126	0.494	0.633	0.952	0.975	1.000	1.000	1.000	1.000	1.000
					Percentile bootstrap					
$\beta_1\beta_2\beta_3 < .01$	0.003	0.008	0.013	0.036	0.088	0.159	0.486	0.692	0.786	0.980
$.01 \le \beta_1\beta_2\beta_3 < .05$	0.042	0.098	0.161	0.243	0.417	0.465	0.852	0.856	0.992	0.992
$.05 \le \beta_1\beta_2\beta_3 > .10$	0.181	0.428	0.598	0.903	0.885	0.999	0.995	1.000	1.000	1.000
$.10 \le \beta_1\beta_2\beta_3$	0.400	0.751	0.808	0.973	0.984	1.000	1.000	1.000	1.000	1.000
					Bias-corrected bootstrap					
$.10 \le \beta_1\beta_2\beta_3 < .01$	0.017	0.036	0.060	0.105	0.215	0.307	0.628	0.818	0.843	0.989
$.01 \le .01\beta_1\beta_2\beta_3 < .05$	0.118	0.190	0.285	0.336	0.538	0.533	0.888	0.879	0.994	0.993
$.05 \le \beta_1\beta_2\beta_3 < .10$	0.359	0.626	0.726	0.948	0.917	0.999	0.996	1.000	1.000	1.000
$.10 \le \beta_1\beta_2\beta_3$	0.566	0.845	0.870	0.982	0.989	1.000	1.000	1.000	1.000	1.000

Note: Table entries are mean power levels, collapsing across study factors for which results were similar. For each sample size, entries in the left column are mean power levels for the dichotomous X condition. Entries in the right column are mean power levels for the normally distributed X condition. Study factors collapsed because of similar results are β4, β5, and β6. The 14 levels of β1, β2, and β3 in which the null hypothesis is false are also collapsed into the 4 levels listed for each method. Results for the product-of-coefficients methods are collapsed, including for the unbiased standard error-only replications in which its variance estimate was nonnegative.

Table 9
Coverage Model Summary, Nontrivial Effects ($R_L^2 \geq .005$)

Effect	LR χ^2	*df*	*p*	R_L^2
Method	48,262.7	4	<.001	.006
β_2	70,427.9	1	<.001	.008
β_3	70,121.4	1	<.001	.008
All other effects ($R_L^2 < .005$)	238,473.9	1273	<.001	.029
Model	391,633.9	1279	<.001	.047

Note: The sum of individual effects' R_L^2 differs from the overall model R_L^2 because R_L^2 is not a true proportion of variance accounted for measure. Results for the unbiased standard error are based on only replications in which its variance estimate was nonnegative. LR = likelihood ratio.

Discussion

This article has introduced three estimators of the standard error of the three-path mediated effect and has compared their performance to three other methods of testing for mediation. The best performers for null hypothesis testing, based on an assessment of Type I error and power, were the joint significance test, the percentile bootstrap, and the bias-corrected bootstrap. The joint significance test and the percentile bootstrap were the more conservative methods: They successfully controlled Type I error and had good power. The bias-corrected bootstrap had consistently the highest power of any tested method, but its Type I error rates were significantly above the nominal level in conditions of large sample size and large nonzero coefficients in the mediated effect.

In terms of coverage performance, the bootstrap methods were the best performers. The percentile bootstrap had too low coverage for the smallest sample size in null false conditions but had coverage near the nominal level in all other conditions. Mirroring its too-high Type I error, the bias-corrected bootstrap had too low coverage in some null hypothesis true conditions, but it performed well in null hypothesis false conditions with sample sizes of at least 200. Of the product-of-coefficients methods, the exact standard error performed best: It almost always had coverage at or above the nominal level, whereas the other two methods had far too low coverage in null hypothesis false conditions with smaller samples.

Of the three product-of-coefficients methods, the multivariate delta estimator performed best in relative bias. It consistently overestimated standard errors by 20% to 30% when two or three of the paths in the mediated effect were zero but for most conditions had negligible relative bias. The unbiased standard error frequently yielded negative variance estimates for the mediated effect in the smallest effect size conditions, and the exact standard error had bias similar to that of the multivariate delta method, but larger.

In summary, we recommend three methods for testing the three-path mediated effect. These are the joint significance test, the percentile bootstrap, and the bias-corrected bootstrap. The major advantage of the joint significance test is its ease and speed of application. In circumstances where only a test of the null hypothesis of no mediation is of interest, it is an ideal method, as it controlled Type I error at or below its nominal level and had good power. Its major drawback is that it cannot be easily used to estimate a confidence interval for the mediated effect. The bootstrap methods allow for confidence

Table 10
Coverage Levels in Null Hypothesis False Conditions and Proportion of Conditions in Which Coverage Was Significantly Below .95

	Sample Size									
	$n = 50$		$n = 100$		$n = 200$		$n = 500$		$n = 1{,}000$	
$\beta_1\beta_2\beta_3$	Coverage	(sig. < .95)	Coverage	(sig. < .95)	Coverage	(sig. < .95)	Coverage	(sig. < .95)	Coverage	(sig. < .95)
					Multivariate delta standard error					
$\beta_1\beta_2\beta_3 < .01$	0.922	(0.734)	0.882	(1.000)	0.880	(1.000)	0.909	(0.984)	0.927	(0.813)
$.01 \leq \beta_1\beta_2\beta_3 < .05$	0.921	(0.759)	0.920	(0.710)	0.930	(0.513)	0.940	(0.290)	0.945	(0.125)
$.05 \leq \beta_1\beta_2\beta_3 < .10$	0.889	(1.000)	0.916	(0.969)	0.932	(0.625)	0.941	(0.234)	0.945	(0.141)
$.10 \leq \beta_1\beta_2\beta_3$	0.909	(1.000)	0.926	(0.844)	0.937	(0.344)	0.947	(0.042)	0.948	(0.063)
					Unbiased standard error					
$\beta_1\beta_2\beta_3 < .01$	0.885	(0.859)	0.839	(1.000)	0.850	(1.000)	0.891	(0.984)	0.919	(0.891)
$.01 \leq \beta_1\beta_2\beta_3 < .05$	0.889	(0.951)	0.900	(0.844)	0.918	(0.683)	0.936	(0.393)	0.943	(0.188)
$.05 \leq \beta_1\beta_2\beta_3 < .10$	0.861	(1.000)	0.902	(0.984)	0.927	(0.859)	0.940	(0.281)	0.944	(0.141)
$.10 \leq \beta_1\beta_2\beta_3$	0.893	(1.000)	0.918	(0.969)	0.934	(0.531)	0.945	(0.052)	0.947	(0.063)
					Exact standard error					
$\beta_1\beta_2\beta_3 < .01$	1.000	(0.000)	0.983	(0.000)	0.938	(0.625)	0.927	(0.781)	0.934	(0.547)
$.01 \leq \beta_1\beta_2\beta_3 < .05$	0.977	(0.000)	0.951	(0.067)	0.942	(0.268)	0.944	(0.147)	0.947	(0.076)
$.05 \leq \beta_1\beta_2\beta_3 < .10$	0.920	(0.859)	0.927	(0.813)	0.936	(0.406)	0.943	(0.156)	0.946	(0.141)
$.10 \leq \beta_1\beta_2\beta_3$	0.925	(0.844)	0.932	(0.594)	0.940	(0.260)	0.948	(0.042)	0.948	(0.063)
					Percentile bootstrap					
$\beta_1\beta_2\beta_3 < .01$	0.994	(0.000)	0.970	(0.000)	0.941	(0.344)	0.944	(0.125)	0.946	(0.109)
$.01 \leq \beta_1\beta_2\beta_3 < .05$	0.950	(0.201)	0.941	(0.241)	0.943	(0.121)	0.946	(0.054)	0.948	(0.031)
$.05 \leq \beta_1\beta_2\beta_3 < .10$	0.930	(0.719)	0.940	(0.234)	0.944	(0.078)	0.946	(0.031)	0.947	(0.047)
$.10 \leq \beta_1\beta_2\beta_3$	0.935	(0.417)	0.940	(0.198)	0.944	(0.104)	0.949	(0.021)	0.949	(0.031)

(continued)

Table 10 (continued)

	Sample Size									
	$n = 50$		$n = 100$		$n = 200$		$n = 500$		$n = 1{,}000$	
$\beta_1\beta_2\beta_3$	Coverage	(sig. < .95)	Coverage	(sig. < .95)	Coverage	(sig. < .95)	Coverage	(sig. < .95)	Coverage	(sig. < .95)
					Bias-corrected bootstrap					
$\beta_1\beta_2\beta_3 < .01$	0.933	(0.547)	0.897	(1.000)	0.927	(0.422)	0.957	(0.000)	0.955	(0.000)
$.01 \leq \beta_1\beta_2\beta_3 < .05$	0.899	(0.987)	0.924	(0.848)	0.943	(0.143)	0.948	(0.063)	0.949	(0.027)
$.05 \leq \beta_1\beta_2\beta_3 < .10$	0.944	(0.250)	0.953	(0.016)	0.953	(0.000)	0.950	(0.000)	0.948	(0.047)
$.10 \leq \beta_1\beta_2\beta_3$	0.949	(0.021)	0.950	(0.031)	0.948	(0.010)	0.950	(0.021)	0.950	(0.031)

Note: Only null hypothesis false conditions are shown because coverage in null hypothesis true conditions = 1 − Type I error rate; coverage levels for these conditions may be inferred from Table 6. For each sample size, entries in the left column are mean coverage levels, collapsing across study factors for which results were similar. Entries in the right column (in parentheses) are the proportions of the collapsed conditions in which the coverage level fell significantly below .95. Each coverage level was tested by forming a 95% confidence interval. As each condition consisted of 1,000 replications, a coverage level of .934 (which has a 95% confidence interval of [.919, .949]) or less fell significantly below .95. Study factors collapsed because of similar results are *X* distribution, β_4, β_5, and β_6. The 14 levels of β_1, β_2, and β_3 in which the null hypothesis is false are also collapsed into the 4 levels listed for each method. Results for the product-of-coefficients methods are collapsed, including for the unbiased standard error- only replications in which its variance estimate was nonnegative.

intervals to be estimated as well as performing hypothesis tests. Of the two, the percentile bootstrap is the more conservative. It controlled Type I error at or below its nominal level in all conditions. The bias-corrected bootstrap had the highest power of any method across conditions but had excess Type I error in some conditions. Although statistically significant, these excess Type I error rates were only .087 at the largest, though, so the extra power the method affords might outweigh the excess Type I error.

An SAS macro that performs all the tests for mediation discussed in this article is available at the second author's Web site (http://www.public.asu.edu/~davidpm/ripl/mediate.htm). The macro works only for manifest variable models. For models that include latent variables, the joint significance test may be easily performed using any structural equation modeling (SEM) program that outputs estimates and standard errors for b_1, b_2, and b_3. These include Mplus (Muthén & Muthén, 2006), LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1995), and AMOS (SPSS Inc., 2006). Confidence intervals for both bootstrap methods may also be obtained from Mplus (using the "model indirect" command and the "cinterval(bootstrap)" or "cinterval(bcbootstrap)" output options). LISREL, EQS, and AMOS will perform bootstrapping and save out bootstrap estimates of the b_1, b_2, and b_3 but will not provide limits of a confidence interval based on these bootstrap estimates. Both percentile and bias-corrected confidence limits may be found from these output values, either manually (e.g., by picking the 2.5th and 97.5th percentiles for a percentile bootstrap 95% confidence interval) or using a second SAS macro also available at the second author's Web site.

Generalizing the Results

The results of the present study are similar to those of previous studies of the two-path mediated effect. MacKinnon and colleagues (2002) also found good Type I error and power performance for the joint significance test. MacKinnon et al. (2004) found the bias-corrected bootstrap to have the most power of any tested method and recommended it based on this result. They also found that it had excess Type I error, though, and that the percentile bootstrap had better control of Type I error (pp. 117, 120), although they considered the extra power to be worth risking a little excess Type I error. Although our recommendation is more conservative in suggesting the percentile bootstrap along with the bias corrected bootstrap, the similarity of results across two- and three-path mediated effects suggests that these best performing methods may be generalized to testing mediational chains of any length.

The Monte Carlo study reported here made several simplifying assumptions, such as that there were no other variables in the model other than the independent variable, the dependent variable, and the two mediators. This condition need not hold for mediation to be tested. The best performing methods require only estimates of the three-path coefficients b_1, b_2, b_3, and their standard errors. The inclusion of other variables in the model should only improve these estimates, and therefore the tests of the mediated effect, particularly when exclusion of the other variables would have resulted in bias in the estimates of the path coefficients.

Another simplifying assumption of the Monte Carlo study was that the variables were all measured without error. In practice, of course, this is a highly unrealistic assumption. Measurement error in the variables in the model will attenuate their zero-order correlations

toward zero. This typically means that the coefficients b_1, b_2, and b_3 will also be attenuated toward zero, although in the models including more than one predictor (Equations 2 and 3), coefficients may be biased in either direction (Cohen, Cohen, West, & Aiken, 2003, Box 4.3.1). Standard errors for the coefficients will increase with measurement error in the dependent variable of each regression (Cohen et al., 2003, p. 124). For both of these reasons, the power results reported here are likely upper bounds on the power that might be expected in models having effect sizes used in the Monte Carlo study. The problem of measurement error in the variables can be dealt with using an SEM. Latent variables can be used in place of any or all of the measured variables X, M_1, M_2, and Y in Figure 1. As the path coefficients b_1, b_2, b_3, and their standard errors can still be estimated, the methods discussed here can all be applied.

In the Baron and Kenny (1986) framework, partial and full mediation are distinguished. Partial mediation occurs when a mediator accounts for only some of the relation between the independent and dependent variables. Full mediation occurs when no significant relation remains between the independent and dependent variables after the mediator is entered in the model. Whether a mediator (or series of mediators as in the three-path situation) partially or fully mediates the independent-to-dependent variable relation can be found by examining the size of the direct effect, which is b_4 in the three-path situation. The size of the corresponding true value β_4 was a factor manipulated in the Monte Carlo study; it did not have a nontrivial effect, either alone or in an interaction term, on any of the study outcomes. This result suggests that the conclusions of the present study may be generalized across instances of both partial and full mediation.

Mediational effects may also be categorized as consistent or inconsistent (MacKinnon, Krull, & Lockwood, 2000). In the single-mediator context, mediation is consistent when the mediated effect has the same sign as the direct effect. Mediation is inconsistent when the mediated effect has the opposite sign as the direct effect. The issue is more complex in the two-mediator case because there are effects that pass through only one mediator ($\beta_1\beta_6$ and $\beta_5\beta_3$) in addition to the mediated effect $\beta_1\beta_2\beta_3$ and the direct effect β_4, and these effects may have any combination of positive and negative signs. In the Monte Carlo study reported here, all coefficients had only positive or zero values. The performance of the methods compared should be unaffected by the signs of the effects, though. The signs of the b_4, b_5, and b_6 paths are irrelevant because these paths are not included in any of the tests; the paths are important only because their estimation allows for better estimates of b_1, b_2, and b_3 and their standard errors. As for the signs of b_1, b_2, and b_3, the joint significance test applies two-tailed tests to each, so its performance will not depend on their signs. Similarly for the product-of-coefficients methods, the point estimate of the mediated effect $b_1b_2b_3$, which is where confidence intervals are centered, will be an unbiased estimate of the true mediated effect $\beta_1\beta_2\beta_3$ as long as each coefficient is an unbiased estimate of its corresponding true value, regardless of the sign on any of the coefficients or true values. The estimates of the standard error of $b_1b_2b_3$ will also not be affected by the signs of the coefficients, as the coefficients are squared when entered in the formulas (see Table 1). Finally, confidence intervals generated using bootstrap methods will be unaffected by the signs of the coefficients. If at least one coefficient has a true value of zero (i.e., the null hypothesis is true), then bootstrap estimates of that coefficient will tend to change sign from one bootstrap sample to another, making the bulk of the bootstrap distribution fall

near zero and leading to a failure to reject the null hypothesis. If all three coefficients are nonzero (i.e., the null hypothesis is false), then bootstrap estimates of the coefficients will tend to have the same sign from one bootstrap sample to another, moving the bulk of the bootstrap distribution away from zero (with the direction depending on the signs of all three coefficients) and leading to a rejection of the null hypothesis.

Limitations and Future Directions

Estimating a three-path mediational model using Equations 1-3 requires making a number of assumptions, mostly relating to the correct specification of the model. As outlined by James, Mulaik, and Brett (1982), correct specification includes specification of the causal order of the variables, specification of the causal direction (no reciprocal paths are estimated), the assumption that the model is self-contained (i.e., there are no omitted variables; this includes the assumption that interactions need not be included), the assumption that there are no moderator effects, and the assumption that the model is stable. As with a simple regression model, the misspecification of the model by failing to include any of these relations or variables when they should be included will lead to bias in the regression coefficients. The three models in Equations 1-3 are also assumed to have independent residuals. This assumption can be tested if the three-path mediation test occurs in the context of a larger SEM, which could allow for the covariance among the residuals to be estimated to discover whether it is, in fact, different from zero.

In addition, as the models that make up the three-path mediational model are regression models, whether estimated using regression or SEM, they make all the assumptions typically made by regression models. These include the assumptions of linearity of the relations between variables, normally distributed residuals when conditioning on the predictors, and residuals that are independent of one another. If the linearity assumption is violated, this is a misspecification of the model, and regression coefficients will tend to be biased. If the assumptions about the residuals are violated, the standard errors will be biased.

The methods used to obtain point estimates and confidence intervals for the three-path mediated effect investigated in this article assume that the estimate of the relation between the mediators and the relation from the mediator to the dependent variable represent the true causal effect. As outlined by Holland (1988), these assumptions may be unlikely to be true. Following Holland's analysis, it is likely that the unadjusted total effect of the treatment on the first mediator, the treatment effect on the second mediator, and the treatment effect on the dependent variable are estimates of the underlying causal effect. The regression coefficients for the relations among the two mediators and the dependent variable are not direct estimates of causal effects in Rubin's (1974) causal model. For example, the ordering of the two mediators and the dependent variable may not be as hypothesized. Researchers should keep in mind that the methods described here make this assumption, which may be wrong for some data. Theory and additional experimental studies must be used to bolster the causal hypothesis underlying the three-path mediation model.

One approach to testing for mediation that has not been mentioned here is the asymmetric distribution of the product method. MacKinnon et al. (2004) applied this method in the single mediator case, where the distribution used is of the product of two random variables, one for each of the two paths. In the three-path situation studied here, the

distribution of the product of three random variables would be required. A few studies provide analytic formulas for this distribution, but no software to integrate the density has yet been introduced. This approach and Monte Carlo approaches to estimate this distribution are a topic for future study.

Appendix

The multivariate delta method is a general method of estimating the variance of functions of random variables that are normally distributed. As the regression coefficients b_1, b_2, and b_3 are normally distributed, this method can be applied to estimate the variance of $b_1b_2b_3$. The two parts required are the covariance matrix of b_1, b_2, and b_3, labeled $\mathbf{V}$, and the vector of partial derivatives of $b_1b_2b_3$ with respect to each of b_1, b_2, and b_3, labeled $\mathbf{d}$. The covariance matrix is pre- and post-multiplied by the vector of derivatives:

$$
\begin{aligned}
s^2_{\text{multivariate delta}} = \mathbf{dVd}' &= \left[\frac{\partial b_1b_2b_3}{\partial b_1}, \frac{\partial b_1b_2b_3}{\partial b_2}, \frac{\partial b_1b_2b_3}{\partial b_3}\right]
\begin{bmatrix} s^2_{b1} & s_{b1b2} & s_{b1b3} \\ s_{b2b1} & s^2_{b2} & s_{b2b3} \\ s_{b3b1} & s_{b3b2} & s^2_{b3} \end{bmatrix}
\begin{bmatrix} \frac{\partial b_1b_2b_3}{\partial b_1} \\ \frac{\partial b_1b_2b_3}{\partial b_2} \\ \frac{\partial b_1b_2b_3}{\partial b_3} \end{bmatrix} \\
&= [b_2b_3,\ b_1b_3,\ b_1b_2]
\begin{bmatrix} s^2_{b1} & s_{b1b2} & s_{b1b3} \\ s_{b2b1} & s^2_{b2} & s_{b2b3} \\ s_{b3b1} & s_{b3b2} & s^2_{b3} \end{bmatrix}
\begin{bmatrix} b_2b_3 \\ b_1b_3 \\ b_1b_2 \end{bmatrix} \\
&= b_2^2b_3^2s^2_{b1} + b_1^2b_3^2s^2_{b2} + b_1^2b_2^2s^2_{b3} + 2b_1b_2b_3^2s_{b1b2} + 2b_1b_2^2b_3s_{b1b3} + 2b_1^2b_2b_3s_{b2b3}.
\end{aligned}
$$

The three-path coefficients b_1, b_2, and b_3 are independent, so the last three terms are zero, and the multivariate delta estimate of the variance is

$$
s^2_{\text{multivariate delta}} = b_1^2b_2^2s^2_{b^3} + b_1^2b_3^2s^2_{b^2} + b_2^2b_3^2s^2_{b^1}.
$$

The unbiased estimate of the variance of $b_1b_2b_3$ is based on the work of Goodman (1960). His Equation 5 gives the unbiased estimate of the variance of two independent random variables, but he suggested that it could be easily extended to more than two variables. Extending his equation to three independent variables, which in this case are the path coefficients b_1, b_2, and b_3, yields

$$
\begin{aligned}
s^2_{\text{unbiased}} &= (b_1^2 - s^2_{b1})(b_2^2 - s^2_{b2})s^2_{b3} + (b_1^2 - s^2_{b1})(b_3^2 - s^2_{b3})s^2_{b2} + (b_2^2 - s^2_{b2})(b_3^2 - s^2_{b3})s^2_{b1} \\
&\quad + (b_1^2 - s^2_{b1})s^2_{b2}s^2_{b3} + (b_2^2 - s^2_{b2})s^2_{b1}s^2_{b3} + (b_3^2 - s^2_{b3})s^2_{b1}s^2_{b2} + s^2_{b1}s^2_{b2}s^2_{b3} \\
&= (b_1^2b_2^2s^2_{b3} - b_1^2s^2_{b2}s^2_{b3} - b_2^2s^2_{b1}s^2_{b3} + s^2_{b1}s^2_{b2}s^2_{b3}) + (b_1^2b_3^2s^2_{b2} - b_1^2s^2_{b2}s^2_{b3} - b_3^2s^2_{b1}s^2_{b2} + s^2_{b1}s^2_{b2}s^2_{b3}) \\
&\quad + (b_2^2b_3^2s^2_{b1} - b_2^2s^2_{b1}s^2_{b3} - b_3^2s^2_{b1}s^2_{b2} + s^2_{b1}s^2_{b2}s^2_{b3}) + (b_1^2s^2_{b2}s^2_{b3} - s^2_{b1}s^2_{b2}s^2_{b3}) \\
&\quad + (b_2^2s^2_{b1}s^2_{b3} - s^2_{b1}s^2_{b2}s^2_{b3}) + (b_3^2s^2_{b1}s^2_{b2} - s^2_{b1}s^2_{b2}s^2_{b3}) + s^2_{b1}s^2_{b2}s^2_{b3} \\
&= b_1^2b_2^2s^2_{b3} + b_1^2b_3^2s^2_{b2} + b_2^2b_3^2s^2_{b1} - b_1^2s^2_{b2}s^2_{b3} - b_2^2s^2_{b1}s^2_{b3} - b_3^2s^2_{b1}s^2_{b2} + s^2_{b1}s^2_{b2}s^2_{b3}.
\end{aligned}
$$

The exact variance estimate is also an extension of Goodman's (1960) exact variance estimate for the product of two random variables. This formula requires the square of the coefficient of variation to be defined. For variable b_i, it is $G(b_i) = s^2_{bi}/b_i^2$. Extending Goodman's Equation 2, the exact variance estimate of $b_1b_2b_3$ is

$$
\begin{aligned}
s^2_{\text{exact}} &= (b_1b_2b_3)^2[G(b_1) + G(b_2) + G(b_3) + G(b_1)G(b_2)G(b_1) + G(b_3) + G(b_2)G(b_3) \\
&\quad + G(b_1)G(b_2)G(b_3)] \\
&= (b_1b_2b_3)^2[(s^2_{b1}/b^2_1) + (s^2_{b2}/b^2_2) + (s^2_{b3}/b^2_3) + (s^2_{b1}/b^2_1)(s^2_{b2}/b^2_2) + (s^2_{b1}/b^2_1)(s^2_{b3}/b^2_3) \\
&\quad + (s^2_{b2}/b^2_2)(s^2_{b3}/b^2_3) + (s^2_{b1}/b^2_1)(s^2_{b2}/b^2_2)(s^2_{b3}/b^2_3)] \\
&= b^2_1b^2_2s^2_{b3} + b^2_1b^2_3s^2_{b2} + b^2_2b^2_3s^2_{b1} + b^2_1s^2_{b2}s^2_{b3} + b^2_2s^2_{b1}s^2_{b3} + b^2_3s^2_{b1}s^2_{b2} + s^2_{b1}s^2_{b2}s^2_{b3}.
\end{aligned}
$$

References

Aiken, L. S., Gerend, M. A., & Jackson, K. M. (2001). Subjective risk and health protective behavior: Cancer screening and cancer prevention. In A. Baum, T. Revenson, & J. Singer (Eds). *Handbook of Health Psychology* (pp. 727-746). Mahwah, NJ: Erlbaum.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.

Allen, D. G., & Griffeth, R. W. (2001). Test of a mediated performance-turnover relationship highlighting the moderating roles of visibility and reward contingency. *Journal of Applied Psychology*, *86*, 1014-1021.

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, *40*, 37-47.

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.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. In C. C. Clogg (Ed.), *Sociological methodology, 1990* (pp. 115-140). Oxford, UK: Basil Blackwell.

Claessens, B. J. C., Van Eerde, W., Rutte, C. G., & Roe, R. A. (2004). Planning behavior and perceived control of time at work. *Journal of Organizational Behavior*, *25*, 937-950.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Craig, C. C. (1936). On the frequency function of *xy*. *Annals of Mathematical Statistics*, *7*, 1-15.

Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). *Socioeconomic background and achievement*. New York: Seminar Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, *55*, 708-713.

Hagle, T. M., & Mitchell, G. E. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, *36*, 762-784.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 449-484). Washington, DC: American Sociological Association.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, *69*, 307-321.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.

James, L. R., Mulaik, S. A., & Brett, J. M. (2006). A tale of two methods. *Organizational Research Methods*, *9*, 233-244.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Lincolnwood, IL: Scientific Software International.

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

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*, 69-86.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 233-265). New York: McGraw-Hill.

Kiefer, T. (2005). Feeling bad: Antecedents and consequences of negative emotions in ongoing change. *Journal of Organizational Behavior*, *26*, 875-897.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*, 144-158.

MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (in press). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavioral Research Methods*.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science*, *1*, 173-181.

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*, 99-128.

Manly, B. F. (1997). *Randomization, bootstrap, and Monte Carlo methods in biology* (2nd ed.). New York: Chapman and Hall.

McGuire, W. J. (1980). The communication-persuasion model and health-risk labeling. In L. A. Morris, M. B. Mazis, & I. Barofsky (Eds.), *Product labeling and health risks* (Banbury Report No. 6, pp. 99-122). New York: Cold Spring Harbor Laboratory.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*, 17-24.

Muthén, L. K., & Muthén, B. (2006). *Mplus user's guide* (Version 4). Los Angeles: Author.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688-701.

SAS Institute. (2005). SAS [Computer software]. Cary, NC: Author.

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, 1982* (pp. 290-312). Washington, DC: American Sociological Association.

Springer, M. D., & Thompson, W. E. (1970). The distribution of products of beta, gamma, and Gaussian random variables. *SIAM Journal on Applied Mathematics*, *18*, 721-737.

SPSS Inc. (2006). AMOS 7.0 [Computer software]. Chicago: Author.

Stewart, G. L., & Barrick, M. R. (2000). Team structure and performance: Assessing the mediating role of intrateam process and the moderating role of task type. *Academy of Management Journal*, *43*, 135-148.

Tein, J.-Y., Sandler, I. N., & Zautra, A. J. (2000). Stressful life events, psychological distress, coping, and parenting of divorced mothers: A longitudinal study. *Journal of Family Psychology*, *14*, 27-41.

Tekleab, A. G., Bartol, K. M., & Liu, W. (2005). Is it pay levels or pay raises that matter to fairness and turnover? *Journal of Organizational Behavior*, *26*, 899-921.

Aaron B. Taylor is a PhD candidate in quantitative psychology at Arizona State University. His research considers resampling methods and structural equation modeling.

David P. MacKinnon is a professor in the Department of Psychology at Arizona State University. His primary interest is in statistical methods to assess the effects of health promotion and disease prevention interventions, especially methods to determine how interventions achieve their effects.

Jenn-Yun Tein, PhD, is currently the codirector of the Research Methodology Core of the Prevention Research Center at Arizona State University. Her research interests are statistical and methodological applications in prevention research. Her areas of specialties include survival analyses, multilevel analyses, mediation/moderation analyses, and structural equation modeling.