Published in final edited form as:
Psychol Methods. 2002 March ; 7(1): 83.

A Comparison of Methods to Test Mediation and Other Intervening Variable Effects

David P. MacKinnon, **Chondra M. Lockwood**, **Jeanne M. Hoffman**, **Stephen G. West**, and **Virgil Sheets**
Department of Psychology, Arizona State University.

Abstract

A Monte Carlo study compared 14 methods to test the statistical significance of the intervening variable effect. An intervening variable (mediator) transmits the effect of an independent variable to a dependent variable. The commonly used R. M. Baron and D. A. Kenny (1986) approach has low statistical power. Two methods based on the distribution of the product and 2 difference-in-coefficients methods have the most accurate Type I error rates and greatest statistical power except in 1 important case in which Type I error rates are too high. The best balance of Type I error and statistical power across all cases is the test of the joint significance of the two effects comprising the intervening variable effect.

The purpose of this article is to compare statistical methods used to test a model in which an independent variable (*X*) causes an intervening variable (*I*), which in turn causes the dependent variable (*Y*). Many different disciplines use such models, with the terminology, assumptions, and statistical tests only partially overlapping among them. In psychology, the $X \rightarrow I \rightarrow Y$ relation is often termed *mediation* (Baron & Kenny, 1986), sociology originally popularized the term *indirect effect* (Alwin & Hauser, 1975), and in epidemiology, it is termed the *surrogate* or *intermediate endpoint effect* (Freedman & Schatzkin, 1992). This article focuses on the statistical performance of each of the available tests of the effect of an intervening variable. Consideration of conceptual issues related to the definition of intervening variable effects is deferred to the final section of the Discussion.

Hypotheses articulating measurable processes that intervene between the independent and dependent variables have long been proposed in psychology (e.g., MacCorquodale & Meehl, 1948; Woodworth, 1928). Such hypotheses are fundamental to theory in many areas of basic and applied psychology (Baron & Kenny, 1986; James & Brett, 1984). Reflecting this importance, a search of the *Social Science Citation Index* turned up more than 2,000 citations of the Baron and Kenny article that presented an important statistical approach to the investigation of these processes. Examples of hypotheses and models that involve intervening variables abound. In basic social psychology, intentions are thought to mediate the relation between attitude and behavior (Ajzen & Fish-bein, 1980). In cognitive psychology, attentional processes are thought to intervene between stimulus and behavior (Stacy, Leigh, & Weingardt, 1994). In industrial psychology, work environment leads to changes in the intervening variable of job perception, which in turn affects behavioral outcomes (James & Brett, 1984). In applied

Correspondence concerning this article should be addressed to David P. MacKinnon, Department of Psychology, Arizona State University, Tempe, Arizona 85287-1104. David.MacKinnon@asu.edu.
Jeanne M. Hoffman is now at the Department of Rehabilitation Medicine, University of Washington. Virgil Sheets is now at the Department of Psychology, Indiana State University.

Editor's Note. Howard Sandler served as the action editor for this article.—SGW

work on preventive health interventions, programs are designed to change proximal variables, which in turn are expected to have beneficial effects on the distal health outcomes of interest (Hansen, 1992; MacKinnon, 1994; West & Aiken, 1997).

A search of psychological abstracts from 1996 to 1999 yielded nearly 200 articles with the term *mediation, mediating,* or *intervening* in the title and many more articles that investigated the effects of intervening variables but did not use this terminology. Both experimental and observational studies were used to investigate processes involving intervening variables. An investigation of a subset of 50 of these articles found that the majority of studies examined one intervening variable at a time. Despite the number of articles proposing to study effects with intervening variables, fewer than a third of the subset included any test of significance of the intervening variable effect. Not surprisingly, the majority of studies that did conduct a formal test used Baron and Kenny's (1986) approach to testing for mediation. One explanation of the failure to test for intervening variable effects in two-thirds of the studies is that the methods are not widely known, particularly outside of social and industrial–organizational psychology. A less plausible explanation is that the large number of alternative methods makes it difficult for researchers to decide which one to use. Another explanation, described below, is that most tests for intervening variable effects have low statistical power.

Our review located 14 different methods from a variety of disciplines that have been proposed to test path models involving intervening variables (see Table 1). Reflecting their diverse disciplinary origins, the procedures vary in their conceptual basis, the null hypothesis being tested, their assumptions, and statistical methods of estimation. This diversity of methods also indicates that there is no firm consensus across disciplines as to the definition of an intervening variable effect. Nonetheless, for convenience, these methods can be broadly conceptualized as reflecting three different general approaches. The first general approach, the causal steps approach, specifies a series of tests of links in a causal chain. This approach can be traced to the seminal work of Judd and Kenny (1981a,1981b) and Baron and Kenny (1986) and is the most commonly used approach in the psychological literature. The second general approach has developed independently in several disciplines and is based on the difference in coefficients such as the difference between a regression coefficient before and after adjustment for the intervening variable (e.g., Freedman & Schatzkin, 1992;McGuigan & Langholtz, 1988;Olkin & Finn, 1995). The difference in coefficients procedures are particularly diverse, with some testing hypotheses about intervening variables that diverge in major respects from what psychologists have traditionally conceptualized as mediation. The third general approach has its origins in sociology and is based on the product of coefficients involving paths in a path model (i.e., the indirect effect; Alwin & Hauser, 1975;Bollen, 1987;Fox, 1980;Sobel, 1982, 1988). In this article, we reserve the term *mediational model* to refer to the causal steps approach of Judd and Kenny (1981a,1981b) and Baron and Kenny (1986) and use the term *intervening variable* to refer to the entire set of 14 approaches that have been proposed.

To date, there have been three statistical simulation studies of the accuracy of the standard errors of intervening variable effects that examined some of the variants of the product of coefficients and one difference in coefficients approach (MacKinnon & Dwyer, 1993; MacKinnon, Warsi, & Dwyer, 1995; Stone & Sobel, 1990). Two other studies included a small simulation to check standard error formulas for specific methods (Allison, 1995; Bobko & Rieck, 1980). No published study to date has compared all of the available methods within the three general approaches. In this article, we determine the empirical Type I error rates and statistical power for all available methods of testing effects involving intervening variables. Knowledge of Type I error rates and statistical power are critical for accurate application of any statistical test. A method with low statistical power will often fail to detect real effects that exist in the population. A method with Type I error rates that exceed nominal rates (e.g., larger than 5% for nominal alpha = .05) risks finding nonexistent effects. The variety of statistical

tests of effects involving intervening variables suggests that they may differ substantially in terms of statistical power and Type I error rates. We first describe methods used to obtain point estimates of the intervening variable effect and then describe standard errors and tests of significance. The accuracy of the methods is then compared in a simulation study.

The Basic Intervening Variable Model

The equations used to estimate the basic intervening variable model are shown in Equations 1, 2, and 3 and are depicted as a path model in Figure 1.

$$Y=\beta_{0(1)}+\tau X+\varepsilon_{(1)} \quad (1)$$

$$Y=\beta_{0(2)}+\tau' X+\beta I+\varepsilon_{(2)} \quad (2)$$

$$I=\beta_{0(3)}+\alpha X+\varepsilon_{(3)} \quad (3)$$

In these equations, *X* is the independent variable, *Y* is the dependent variable, and *I* is the intervening variable. $\beta_{0(1)}$, $\beta_{0(2)}$, $\beta_{0(3)}$ are the population regression intercepts in Equations 1, 2, and 3, respectively, τ represents the relation between the independent and dependent variables in Equation 1, τ' represents the relation between the independent and dependent variables adjusted for the effects of the intervening variable in Equation 2, α represents the relation between the independent and intervening variables in Equation 3, β represents the relation between the intervening and the dependent variables adjusted for the effect of the independent variable in Equation 2, and $\in_{(1)}$, $\in_{(2)}$, and $\in_{(3)}$ are the residuals in Equations 1, 2, and 3, respectively. For ease of presentation, we use population parameters in all equations, recognizing that in practice the population values are replaced by unbiased sample-based values (e.g., α by $\widehat{\alpha}$) and that our simulations report sample-based estimates of population parameters. Throughout this article, we assume that continuous *X*, *I*, and *Y* have a multivariate normal distribution and that error terms are normally distributed.

Psychological researchers have traditionally focused on testing one of three forms of hypotheses about intervening variables: (a) a series of tests of the causal steps necessary to establish the conditions for mediation proposed by Judd and Kenny (1981b) and Baron and Kenny (1986); (b) tests of each path involved in the effect (α and β); or (c) a test of the product of the two paths (αβ) from Equations 2 and 3. Although we more fully address some of the similarities and differences between these approaches below, one similarity is of importance here. Some methods (Baron & Kenny, 1986; Judd & Kenny, 1981a, 1981b; McGuigan & Langholtz, 1988) use the difference in the independent variable coefficients (τ − τ') in Equations 1 and 2 to estimate the value of the intervening variable effect. If the independent variable coefficient (τ') does not differ significantly from zero when the intervening variable is included in the model, then the results are consistent with a model in which the effect is completely transmitted through the intervening variable (i.e., H_0: τ' = 0 cannot be rejected, where H_0 is the null hypothesis). MacKinnon et al. (1995) have shown that τ − τ' is algebraically equivalent to αβ for ordinary least-squares regression, so that the null hypothesis for the τ − τ' test and the test of αβ are identical. The methodological advantages and disadvantages of the three forms of the null hypothesis are a matter of some debate that we will revisit in the Discussion. We also note that some of the methods proposed to test intervening variable effects test still other null hypotheses. These will be identified during our presentation of the 14 methods.

Causal Steps Tests of the Intervening Variable Effect

Methods to assess intervening variable effects based on causal steps entail tests of different logical relations among the three variables involved. Each must be true for the basic intervening variable (mediational) model to hold. As shown in Table 1, three variants of the causal steps method that test three slightly different hypotheses have been proposed.

The sequence of causal steps outlined in the work of Judd and Kenny (1981a, 1981b) was originally proposed in the context of probing the causal mediation process through which a treatment produces an outcome. Reflecting this context, Judd and Kenny included statistical tests that can help rule out some alternatives to the hypothesized mediational process $X \rightarrow M \rightarrow Y$ of focal interest, where M is a mediator (intervening variable). Baron and Kenny (1986) more explicitly extended the Judd and Kenny approach to contexts in which the independent variable is also measured. Although the overall purpose of this approach focuses on establishing conditions that Judd and Kenny argue are necessary for mediation to occur, the causal steps approach is used to establish the statistical significance of the intervening variable effect by testing each logical relation.

The series of causal steps described by Judd and Kenny (1981b) and Baron and Kenny (1986) differ only slightly. Judd and Kenny (1981b, p. 605) require three conclusions for mediation: (a) “The treatment affects the outcome variable” (H_0: $\tau = 0$), (b) “Each variable in the causal chain affects the variable that follows it in the chain, when all variables prior to it, including the treatment, are controlled” (H_0: $\alpha = 0$ and H_0: $\beta = 0$), and (c) “The treatment exerts no effect upon the outcome when the mediating variables are controlled” (H_0: $\tau' = 0$). Baron and Kenny (p. 1176) defined three conditions for mediation: (a) “Variations in levels of the independent variable significantly account for variations in the presumed mediator” (H_0: $\alpha = 0$), (b) “Variations in the mediator significantly account for variations in the dependent variable” (H_0: $\beta = 0$), and (c) “When Paths a [α] and b [β] are controlled, a previously significant relation between independent and dependent variables is no longer significant, with the strongest demonstration of mediation occurring when Path c [τ'] is zero” (H_0: $\tau' = 0$). Implicit in condition c is the requirement of an overall significant relation between the independent and dependent variables (H_0: $\tau = 0$). The central difference between the two variants is that Judd and Kenny emphasized the importance of demonstrating complete mediation, which would occur when the hypotheses that $\tau' = 0$ cannot be rejected. Baron and Kenny argued that models in which there is only partial mediation (i.e., $|\tau'| < |\tau|$) rather than complete mediation are acceptable. They pointed out that such models are more realistic in most social science research because a single mediator cannot be expected to completely explain the relation between an independent and a dependent variable.

A third variant of the causal steps approach has also been used by some researchers (Cohen & Cohen, 1983, p. 366). In this variant, researchers claim evidence for intervening variable effects when separate tests of each path in the intervening variable effect (α and β) are jointly significant (H_0: $\alpha = 0$ and H_0: $\beta = 0$). This method simultaneously tests whether the independent variable is related to the intervening variable and whether the intervening variable is related to the dependent variable. A similar test was also outlined by Allison (1995). Kenny, Kashy, and Bolger (1998) restated the Judd and Kenny causal steps but noted that the tests of α and β are the essential tests for establishing mediation. This method provides the most direct test of the simultaneous null hypothesis that path α and path β are both equal to 0. However, this method provides no test of either the $\alpha\beta$ product or the overall $X \rightarrow Y$ relation.

In summary, the widely used Judd and Kenny (1981b) and Baron and Kenny (1986) variants of the causal steps approach clearly specify the conceptual links between each hypothesized causal relation and the statistical tests of these links. However, as described in more depth in

the Discussion, these two approaches probe, but do not provide, the full set of necessary conditions for the strong inference of a causal effect of the independent variable on the dependent variable through the intervening variable, even when subjects are randomly assigned to the levels of the independent variable in a randomized experiment (Baron & Kenny, 1986; Holland, 1988; MacKinnon, 1994). Because the overall purpose of the causal steps methods was to establish conditions for mediation rather than a statistical test of the indirect effect of *X* on *Y* through *I* (e.g., αβ), they have several limitations. The causal steps methods do not provide a joint test of the three conditions (conditions a, b, and c), a direct estimate of the size of the indirect effect of *X* on *Y*, or standard errors to construct confidence limits, although the standard error of the indirect effect of *X* on *Y* is given in the descriptions of the causal steps method (Baron & Kenny, 1986; Kenny et al., 1998). In addition, it is difficult to extend the causal steps method to models incorporating multiple intervening variables and to evaluate each of the intervening variable effects separately in a model with more than one intervening variable (e.g., MacKinnon, 2000; West & Aiken, 1997). Finally, the requirement that there has to be a significant relation between the independent and dependent variables excludes many "inconsistent" intervening variable models in which the indirect effect (αβ) and direct effect (τ') have opposite signs and may cancel out (MacKinnon, Krull, & Lockwood, 2000).

Difference in Coefficients Tests of the Intervening Variable Effect

Intervening variable effects can be assessed by comparing the relation between the independent variable and the dependent variable before and after adjustment for the intervening variable. Several different pairs of coefficients can be compared, including the regression coefficients ($\tau - \tau'$) described above and correlation coefficients, $\rho_{XY} - \rho_{XY.I}$, although, as described below, the method based on correlations differs from other tests of the intervening variable effect. In the above expression, ρ_{XY} is the correlation between the independent variable and the dependent variable and $\rho_{XY.I}$ is the partial correlation between the independent variable and the dependent variable partialled for the intervening variable. Each of the variants of the difference in coefficients tests is summarized in the middle part of Table 1. Readers should note that these procedures test a diverse set of null hypotheses about intervening variables.

Freedman and Schatzkin (1992) developed a method to study binary health measures that can be extended to the difference between the adjusted and unadjusted regression coefficients (H_o: $\tau - \tau' = 0$). Freedman and Schatzkin derived the correlation between τ and τ' that can be used in an equation for the standard error based on the variance and covariance of the adjusted and unadjusted regression coefficients:

$$\sigma_{\text{Freedman–Schatzkin}} = \sqrt{\sigma_{\tau}^2 + \sigma_{\tau'}^2 - 2\sigma_{\tau}\sigma_{\tau'}\sqrt{1-\rho_{XI}^2}}. \quad (4)$$

In this equation, ρ_{XI} is equal to the correlation between the independent variable and the intervening variable, σ_t is the standard error of τ, and $\sigma_{\tau'}$ is the standard error of τ'. The estimate of $\tau - \tau'$ is divided by the standard error in Equation 4 and this value is compared to the *t* distribution for a test of significance.

McGuigan and Langholtz (1988) also derived the standard error of the difference between these two regression coefficients (H_o: $\tau - \tau' = 0$) for standardized variables. They found the standard error of the $\tau - \tau'$ method to be equal to

$$\sigma_{\text{McGuigan–Langholtz}} = \sqrt{\sigma_{\tau}^2 + \sigma_{\tau'}^2 - 2(\rho_{\tau\tau'}\sigma_{\tau}\sigma_{\tau'})}. \quad (5)$$

The covariance between τ and τ' ($\rho_{\tau\tau'}\sigma_\tau\sigma_{\tau'}$), applicable for either standardized or unstandardized variables, is the mean square error (σ_{MSE}) from Equation 2 divided by the product of sample size and the variance of *X*. The difference between the two regression coefficients (τ − τ') is then divided by the standard error in Equation 5 and this value is compared to the *t* distribution for a significance test. MacKinnon et al. (1995) found that the original formula for the standard error proposed by McGuigan and Langholtz was inaccurate for a binary (i.e., unstandardized) independent variable. On the basis of their derivation, we obtained the corrected formula given above that is accurate for standardized or unstandardized independent variables.

Another estimate of the standard error of τ − τ' was developed from a test of collapsibility (Clogg, Petkova, & Shihadeh, 1992). Clogg et al. extended the notion of collapsibility in categorical data analysis to continuous measures. Collapsibility tests whether it is appropriate to ignore or collapse across a third variable when examining the relation between two variables. In this case, collapsibility is a test of whether an intervening variable significantly changes the relation between two variables. As shown below, the standard error of τ − τ' in Clogg et al. (1992) is equal to the absolute value of the correlation between the independent variable and the intervening variable times the standard error of τ':

$$\sigma_{\text{Clogg et al.}} = \frac{\sigma_{MSE}|\rho_{XI}|}{\sqrt{\sigma_X\left[n\left(1-\rho_{XI}^2\right)\right]}} = |\rho_{XI}|\sigma_{\tau'}. \tag{6}$$

Furthermore, Clogg et al. (1992) showed that the statistical test of τ − τ' divided by its standard error is equivalent to testing the null hypothesis, H_o: β = 0. This indicates that a significance test of the intervening variable effect can be obtained simply by testing the significance of β or by dividing τ − τ' by the standard error in Equation 6 and comparing the value to the *t* distribution. Although Clogg et al.'s (1992) approach tests H_o: τ − τ' = 0, Allison (1995) and Clogg, Petkova, and Cheng (1995) demonstrate that the derivation assumes that both *X* and *I* are fixed, which is unlikely for a test of an intervening variable.

A method based on correlations (H_o: $\rho_{XY} - \rho_{XY.I} = 0$) compares the correlation between *X* and *Y* before and after it is adjusted for *I* as shown in Equation 7. The difference between the simple and partial correlations is the measure of how much the intervening variable changes the simple correlation, where ρ_{IY} is the correlation between the intervening variable and the dependent variable. The null hypothesis for this test is distinctly different from any of the three forms that have been used in psychology (see p. 86). As noted by a reviewer, there are situations where the difference between the simple and partial correlation is nonzero, yet the correlation between the intervening variable and the dependent variable partialled for the independent variable is zero. In these situations, the method indicates an intervening variable effect when there is not evidence that the intervening variable is related to the dependent variable. The problem occurs because of constraints on the range of the partial correlation:

$$\rho_{\text{difference}} = \rho_{XY} - \frac{\rho_{XY} - \rho_{IY}\rho_{XI}}{\sqrt{\left(1-\rho_{IY}^2\right)\left(1-\rho_{XI}^2\right)}}. \tag{7}$$

Olkin and Finn (1995) used the multivariate delta method to find the large sample standard error of the difference between a simple correlation and the same correlation partialled for a third variable. The difference between the simple and partial correlation is then divided by the calculated standard error and compared to the standard normal distribution to test for an intervening variable effect. The large sample solutions for the variances and covariances among

the correlations are shown in Appendix A and the vector of partial derivatives shown in Equation 8 is used to find the standard error of the difference. There is a typographical error in this formula in Olkin and Finn (p. 160) that is corrected in Equation 8. The partial derivatives are

$$\left[\frac{\rho_{IY}-\rho_{XI}\rho_{XY}}{\left(1-\rho_{IY}^2\right)^{1/2}\left(1-\rho_{XI}^2\right)^{3/2}}\ ,\ 1-\frac{1}{\sqrt{\left(1-\rho_{IY}^2\right)\left(1-\rho_{XI}^2\right)}},\ \frac{\rho_{XI}-\rho_{XY}\rho_{IY}}{\left(1-\rho_{XI}^2\right)^{1/2}\left(1-\rho_{IY}^2\right)^{3/2}}\right]. \quad (8)$$

For the two tests of intervening variable effects (the other test is described below) where standard errors are derived using the multivariate delta method, the partial derivatives are presented in the text rather than showing the entire formula for the standard error, which is long. A summary of the multivariate delta method is shown in Appendix A. An SAS (Version 6.12) program to compute the standard errors is shown in Appendix B.

In summary, each difference in coefficients method provides an estimate of some intervening variable effect and its standard error. Depending on the procedure, the null hypothesis may or may not resemble ones commonly proposed in psychology. The null hypothesis of the Clogg et al. (1992) test assumes fixed *X* and *I*, which is not likely for intervening variables. The difference between the simple and partial correlation represents a unique test of the intervening variable effect because there are situations where there appears to be no relation between the intervening and dependent variable, yet the method suggests that an intervening variable effect exists. An additional drawback of this general approach is that the underlying model for some tests, such as the difference between simple and partial correlation, is based on nondirectional correlations that do not directly follow but are implied by the path model in Figure 1. The difference in coefficients methods also does not provide a clear framework for generalizing the tests to estimate appropriate coefficients and test significance of their difference in models with more than one intervening variable.

Product of Coefficients Tests for the Intervening Variable Effect

The third general approach is to test the significance of the intervening variable effect by dividing the estimate of the intervening variable effect, αβ, by its standard error and comparing this value to a standard normal distribution. There are several variants of the standard error formula based on different assumptions and order of derivatives in the approximations. These variants are summarized in the bottom part of Table 1.

The most commonly used standard error is the approximate formula derived by Sobel (1982) using the multivariate delta method based on a first order Taylor series approximation:

$$\sigma_{\alpha\beta\quad \text{first}}=\sqrt{\alpha^2\sigma_\beta^2+\beta^2\sigma_\alpha^2}. \quad (9)$$

The intervening variable effect is divided by the standard error in Equation 9, which is then compared to a standard normal distribution to test for significance (H_o: αβ = 0). This standard error formula is used in covariance structure programs such as EQS (Bentler, 1997) and LISREL (Jöreskog & Sörbom, 1993).

The exact standard error based on first and second order Taylor series approximation (Aroian, 1944) of the product of α and β is

$$\sigma_{\alpha\beta \quad \text{second}} = \sqrt{\alpha^2\sigma_\beta^2 + \beta^2\sigma_\alpha^2 + \sigma_\alpha^2\sigma_\beta^2}. \tag{10}$$

The intervening variable effect is divided by the standard error in Equation 10, which is then compared to a standard normal distribution to test for significance (H_o: αβ = 0). Equation 9 excludes the product of the two variances, which is part of the exact standard error in Equation 10, although that term is typically very small.

Goodman (1960; Sampson & Breunig, 1971) derived the unbiased variance of the product of two normal variables, which subtracts the product of variances, giving

$$\sigma_{\alpha\beta \quad \text{unbiased}} = \sqrt{\alpha^2\sigma_\beta^2 + \beta^2\sigma_\alpha^2 - \sigma_\alpha^2\sigma_\beta^2}. \tag{11}$$

A test of the intervening variable effect can be obtained by dividing αβ by the standard error in Equation 11, which is then compared to a standard normal distribution to test for significance.

MacKinnon, Lockwood, and Hoffman (1998) showed evidence that the $\alpha\beta/\sigma_{\alpha\beta}$ methods to test the significance of the intervening variable effect have low power because the distribution of the product of regression coefficients α and β is not normally distributed, but rather is often asymmetric with high kurtosis. Under the conditions of multivariate normality of *X*, *I*, and *Y*, the two paths represented by α and β in Figure 1 are independent (MacKinnon, Warsi, & Dwyer, 1995; Sobel, 1982). On the basis of the statistical theory of the products of random variables (Craig, 1936; Meeker, Cornwell, & Aroian, 1981; Springer & Thompson, 1966), MacKinnon and colleagues (MacKinnon et al., 1998; MacKinnon & Lockwood, 2001) proposed three alternative variants (presented below) that theoretically should be more accurate: (a) empirical distribution of $\alpha\beta/\sigma_{\alpha\beta}$ (H_o: $\alpha\beta/\sigma_{\alpha\beta} = 0$), (b) distribution of the product of two standard normal variables, $z_\alpha z_\beta$ (H_o: $z_\alpha z_\beta = 0$), and (c) asymmetric confidence limits for the distribution of the product, αβ (H_o: αβ = 0).

In the first variant, MacKinnon et al. (1998) conducted extensive simulations to estimate the empirical sampling distribution of αβ for a wide range of values of α and β. On the basis of these empirical sampling distributions, critical values for different significance levels were determined. These tables of critical values are available at http://www.public.asu.edu/~davidpm/ripl/methods.htm. For example, the empirical critical value is .97 for the .05 significance level rather than 1.96 for the standard normal test of αβ = 0. We designate this test statistic by z' because it uses a different distribution than the normal distribution.

A second variant of the test of the intervening variable effect involves the distribution of the product of two *z* statistics—one for the α parameter, $z_\alpha = \alpha/\sigma_\alpha$, and another for the β parameter, $z_\beta = \beta/\sigma_\beta$. If α and β are assumed to be normal, the $z_\alpha z_\beta$ term can be directly tested for significance using critical values based on the theoretical distribution of the product of two normal random variables, $P = z_\alpha z_\beta$. This test involves converting both the α and the β paths to *z* scores, multiplying the *z*s, and using a critical value based on the distribution of the product of random variables, $P = z_\alpha z_\beta$, from Craig (1936; see also Meeker et al., 1981; Springer & Thompson, 1966) to determine significance. For example, the critical value to test αβ = 0 for the .05 significance level for the $P = z_\alpha z_\beta$ distribution is 2.18, rather than 1.96 for the normal distribution.

A third variant constructs asymmetric confidence limits to accommodate the nonnormal distribution of the intervening variable effect based on the distribution of the product of random variables. Again, two *z* statistics are computed, $z_\alpha = \alpha/\sigma_\alpha$ and $z_\beta = \beta/\sigma_\beta$. These values are then used to find critical values for the product of two random variables from the tables in Meeker et al. (1981) to find lower and upper significance levels. Those values are used to compute lower and upper confidence limits using the formula CL = αβ± (critical value) $\sigma_{\alpha\beta}$. If the confidence interval does not include zero, the intervening variable effect is significant.

Bobko and Rieck (1980) examined intervening variable effects in path analysis using regression coefficients from the analysis of standardized variables (H_o: $\alpha_\sigma\beta_\sigma = 0$, where α_σ and β_σ are from regression analysis of standardized variables). These researchers used the multivariate delta method to find an estimate of the variance of the intervening variable effect for standardized variables, based on the product of the correlation between *X* and *I* and the partial regression coefficient relating *I* and *Y*, controlling for *X*. The function of the product of these terms is

$$\rho_{\text{product}} = \frac{\rho_{XI}\left(\rho_{IY} - \rho_{XY}\rho_{XI}\right)}{1 - \rho_{XI}^2}. \tag{12}$$

The partial derivatives of this function given in Bobko and Rieck (1980) are

$$\left[\frac{\rho_{XI}^2\rho_{IY} + \rho_{IY} - 2\rho_{XI}\rho_{XY}}{\left(1 - \rho_{XI}^2\right)^2}, \frac{-\rho_{XI}^2}{1 - \rho_{XI}^2}, \frac{\rho_{XI}}{1 - \rho_{XI}^2}\right]. \tag{13}$$

The variance–covariance matrix of the correlation coefficients is pre- and postmultiplied by the vector of partial derivatives to calculate a standard error that can be used to test the significance of the intervening variable effect.

The product of coefficients methods provide estimates of the intervening variable effect and the standard error of the intervening variable effect. In addition, the underlying model follows directly from path analysis wherein the intervening variable effect is the product of coefficients hypothesized to measure causal relations. This logic directly extends to models incorporating multiple intervening variables (Bollen, 1987). However, as is presented below, two problems occur in conducting these tests. First, the sampling distribution of these tests does not follow the normal distribution as is typically assumed. Second, the form of the null hypothesis that is tested is complex.

Overview of Simulation Study

The purpose of the simulation study was to provide researchers with information about the statistical performance of the 14 tests of the intervening variable effect. The primary focus of our comparison was the Type I error rate and statistical power of each test. Intervening variable effect estimates and standard errors were also examined to provide another indication of the accuracy of the methods. We predicted that the use of multiple hypothesis tests in the causal steps approaches would lead to low Type I error rates and low statistical power. We also predicted that many of the traditional tests of the αβ product would also have low Type I error rates and low statistical power because of the associated highly heavy tailed distribution. A central question addressed by the simulation was whether the alternative and the newer tests of the intervening variable effects would yield higher levels of statistical power without increasing the Type I error rate.

Method

Simulation Description

The SAS (Version 6.12) programing language was used for all statistical simulations and analyses. Variables were generated from the normal distribution using the RANNOR function with the current time as the seed. Sample sizes were chosen to be comparable to those common in the social sciences: 50, 100, 200, 500, and 1,000. Parameter values α, β, and τ' were chosen to correspond to effect sizes of zero, small (2% of the variance in the dependent variable), medium (13% of the variance in the dependent variable), and large (26% of the variance in the dependent variable), as described in Cohen (1988, pp. 412–414). These parameters were 0, 0.14, 0.39, and 0.59, corresponding to partial correlations of 0, 0.14, 0.36, and 0.51, respectively. The intervening variable and dependent variable were always simulated to be continuous. In half of the simulations, the independent variable was continuous and in the other half, the independent variable was binary with an equal number of cases in each category. The binary case was included to investigate intervening variable effects in experimental studies. The α parameters were adjusted in the binary case to maintain the same partial correlations as in the continuous case.

In summary, the simulation used a 2 × 4 × 4 × 4 × 5 factorial design. We varied the factors of independent variable type (continuous and binary), effect size of path α (zero, small, medium, and large), effect size of path β, effect size of path τ', and sample size (50, 100, 200, 500, 1,000), for a total of 640 different conditions. A total of 500 replications of each condition were conducted.

Accuracy of Point Estimates and Standard Error

Bias and relative bias were used to assess the accuracy of the point estimates of the intervening variable effect. As shown below, relative bias was calculated as the ratio of bias (numerator) to the true value:

$$RelativeBias = \frac{\widehat{\omega} - \omega}{\omega}, \tag{14}$$

$\widehat{\omega}$ is the point estimate of the simulated intervening variable effect and ω is the true value of the intervening variable effect.

The accuracy of each standard error was determined by comparing the average estimate of the standard error of the intervening variable effect across the 500 simulations to the standard deviation of the intervening variable effect estimate from the 500 simulations. The standard deviation of the intervening variable effect across 500 simulations was the estimate of the true standard error (Yang & Robertson, 1986).

Calculation of Empirical Power and Type I Error Rate

Empirical power or Type I error rates as appropriate were calculated for each test of the intervening variable effect. We report results for the 5% level of significance because it is the most commonly used value in psychology. For each condition, the proportion of times that the intervening variable effect was statistically significant in 500 replications was tabulated.

When $\alpha = 0$, $\beta = 0$, or both α and $\beta = 0$, the proportion of replications in which the null hypothesis of no intervening variable effect was rejected provided an estimate of the empirical Type I error rate. Because we used the 5% significance level, the intervening variable effect was

expected to be statistically significant in 25 (5%) of the 500 samples when the intervening variable effect equals zero.

When both α and β did not equal zero, the proportion of times that each method led to the conclusion that the intervening variable effect was significant provided the measure of statistical power. The higher the proportion of times a method led to the conclusion to reject the false null hypothesis of no effect, the greater the statistical power.

The 14 procedures reference different statistical distributions. In each case, we used the critical values from the reference distribution for the test. For those tests based on asymptotic methods, we used 1.96 from the normal distribution. For the $z' = \alpha\beta/\sigma_{\alpha\beta}$ test we used critical values described in MacKinnon et al. (1998) indicated by z', and for the test for $P = z_\alpha z_\beta$ we used critical values from Craig (1936) indicated by P. The upper and lower confidence limits for the asymmetric confidence limits test were taken from the tabled values in Meeker et al. (1981). The causal steps tests involve multiple hypothesis tests so there is no single reference distribution. For these tests, the intervening variable effect was considered significant if each of the steps was satisfied.

Results

In general, the simulation results for the binary case did not differ from those of the continuous independent variable case. Consequently, we present only the results for the continuous independent variable below.

Intervening Variable Effect Estimates

As found in other studies, most estimates of the intervening variable effect had minimal bias with the exception of $z_\alpha z_\beta$, which had substantial bias because the point estimates of this quantity were much larger than other point estimates of the intervening variable effect. Only the $z_\alpha z_\beta$ test had bias greater than .01, even at a sample size of 50. Relative bias decreased as sample size and effect size increased for all estimates, including $z_\alpha z_\beta$.

Accuracy of Standard Errors

The accuracy of the formulas for standard errors of the intervening variable effect was examined by comparing a measure of the true standard error (the standard deviation of the intervening variable effect estimate across the 500 replications in each condition) to the average standard error estimate as shown in Table 2. For the $\alpha\beta = \tau - \tau'$ estimate, all standard errors were generally accurate except for the Freedman and Schatzkin (1992) and the Clogg et al. (1992) standard error estimates, which were much smaller than the true values for all conditions. Goodman's (1960) unbiased method frequently yielded undefined (imaginary) standard errors. These findings raise serious issues about the use of this approach. For example, Goodman's unbiased standard error was undefined approximately 40% of the time when the true effect size was zero and 10% of the time when the effect size was small and sample size was 50. We did not include cases that resulted in undefined standard errors in the computation of the mean standard error in Table 2.

As shown in Table 3, the standard errors for the product of regression coefficients for standardized variables and standard errors for simple minus partial correlations were all very close to the true values for all conditions, indicating that the standard errors derived using the multivariate delta method were generally accurate.

Power and Type I Error

To reduce the number of tables, we present the results from the subset of conditions in which $\alpha = \beta$ and $\tau' = 0$ in Tables 4, 5, and 6. The results for conditions having nonzero values of τ' and the conditions in which $\alpha \neq \beta$ but both α and β were greater than zero generally produced the same results. Results that differed across the values of τ' are described in the text. The results for the case when either α or β were zero and the other path was nonzero are shown in Tables 7, 8, and 9. The full results of the simulation are available from the Web site http://www.public.asu.edu/~davidpm/ripl/methods.htm.

The causal steps methods had Type I error rates below nominal values at all sample sizes as shown in Table 4. The Baron and Kenny (1986) and Judd and Kenny (1981b) methods had low power for small and medium effect sizes and attained power of .80 or greater for large effects with more than 100 subjects. The Baron and Kenny (1986) method had greater power as τ' increased and the Judd and Kenny (1981b) method had less power as τ' increased. The test of the joint significance of α and β was similar to the other causal steps methods in that it had low Type I error rates. The Type I error rate was consistent with $.05^2 = .0025$ expected for two independent tests, however. Unlike the Baron and Kenny and Judd and Kenny methods, it had at least .80 power to detect large effects at a sample size of 50, medium effects at 100, and approached .80 power to detect a small effect for a sample size of 500. The power to detect small effects was low for all causal steps methods. The joint significance of α and β was the most powerful of the causal steps methods.

Similar to the causal steps methods, all of the difference in coefficients methods had low Type I error rates with two exceptions, as shown in Table 5. All of the $\tau - \tau'$ methods had .80 or greater power and were able to detect small effects once the sample size reached 1,000, medium effects at 100, and large effects at a sample size of 50. Only the Clogg et al. (1992) and Freedman and Schatzkin (1992) methods had accurate Type I error rates (i.e., close to .05) and greater than .80 power to detect a small, medium, and large effect at sample sizes of 500, 100, and 50, respectively. Even though the standard errors from these methods appeared to underestimate the true standard error, they had the most accurate Type I error rates and higher statistical power. This pattern of results suggests that a standard error that is too small may be partially compensating for the higher critical values associated with the nonnormal distribution of the intervening variable effect.

Like most of the previous methods, the product of coefficients methods generally had Type I error rates below .05 and adequate power to detect small, medium, and large effects for sample sizes of 1,000, 100, and 50 respectively. The distribution of products test, $P = z_\alpha z_\beta$, and the distribution of $z' = \alpha\beta/\sigma_{\alpha\beta}$ test had accurate Type I error rates and the most power of all tests. These results are presented in Table 6. At a sample size of 50, the two distribution methods had power of above .80 to detect medium and large effects and detected small effects with .80 power at a sample size of 500. The asymmetric confidence limits method also had Type I error rates that were too low but had more power than the other product of coefficient methods.

Overall, the two distribution methods, $P = z_\alpha z_\beta$, and $z' = \alpha\beta/\sigma_{\alpha\beta}$, the Clogg et al. (1992), and Freedman and Schatzkin (1992) methods performed the best of all of the methods tested in terms of the most accurate Type I error rates and the greatest statistical power. However, recall that the Clogg et al. (1992) method assumes fixed effects for *X* and *I* (equivalent to a test of the significance of β) so that it may not be a good test on conceptual grounds. The similar performance of the Freedman and Schatzkin test suggests that this method is also based on fixed effects for *X* and *I*. These methods were superior whenever both $\alpha = 0$ and $\beta = 0$ and for all combinations of α and β values as long as both parameters were nonzero. For cases when either $\alpha = 0$ and β was nonzero or a was nonzero and $\beta = 0$, these methods were not the most accurate (see Tables 7, 8, and 9). The a path could be nonsignificant and quite small, yet these

methods would suggest a statistically significant intervening variable effect when the β path was a medium or large effect. In the case where α = 0 and the β effect was large, these methods yielded Type I error rates that were too high, although the distribution methods, $P = z_\alpha z_\beta$ and $z' = \alpha\beta/\sigma_{\alpha\beta}$, performed better than the difference in coefficients methods. When α was a large effect and β = 0, the Clogg et al. (1992) and Freedman and Schatzkin methods worked well and the distribution of products methods did not. The joint significance test of α and β, the asymmetric confidence limits test, and the tests based on dividing the αβ intervening variable effect by the standard error of αβ had more accurate standard errors in the case where one of the α and β parameters was equal to zero.

Discussion

In our discussion, we initially focus on the statistical performance of each of the 14 tests of the effect of the intervening variable effect that were considered. We then focus on statistical recommendations and more general conceptual and practical issues associated with the choice of a test of an intervening variable effect.

Statistical Performance

The most widely used methods proposed by Judd and Kenny (1981b) and Baron and Kenny (1986) have Type I error rates that are too low in all the simulation conditions and have very low power, unless the effect or sample size is large. For example, these methods have only .106 empirical power to detect small effects at a sample size of 1,000 and only .49 power to detect moderate effects at a sample size of 200. Overall, the step requiring a significant total effect of *X* on *Y* (τ) led to the most Type II errors. As a result, the Baron and Kenny (1986) causal steps method had fewer Type II errors as the value of τ' increased. The Judd and Kenny (1981b) causal steps method had more Type II errors as τ' increased because of the requirement that τ' not be statistically significant. Studies that use the causal steps methods described by Kenny and colleagues are the most likely to miss real effects but are very unlikely to commit a Type I error. An alternative causal steps method, the test of whether α and β are jointly statistically significant has substantially more power and more accurate Type I error rates.

The power rates for the difference in coefficients methods tend to be higher than the Baron and Kenny (1986) and the Judd and Kenny (1981b) causal steps methods, but the Type I error rates remain too conservative for all but the Clogg et al. (1992) and Freedman and Schatzkin (1992) tests. Although the standard error for the Clogg et al. (1992) and the Freedman and Schatzkin tests do not appear to give an Freedman and Schatzkin tests do not appear to give an Freedman and Schatzkin tests do not appear to give an accurate estimate of the standard error of the intervening variable effect (because of the assumption of fixed *X* and *I*), the significance tests have the most accurate Type I error rates and the greatest statistical power for most situations. Similarly, the product of coefficients methods have higher power than the Baron and Kenny and Judd and Kenny (1981b) methods but, again, the Type I error rates are too low. The low power rates and low Type I error rates are present for the first-order test used in covariance structure analysis programs including LISREL (Jöreskog & Sörbom, 1993) and EQS (Bentler, 1997). The distribution of the product test $P = z_\alpha z_\beta$ and the distribution of $z' = \alpha\beta/\sigma_{\alpha\beta}$ have accurate Type I error rates when α = β = 0, and the highest power rates throughout. These two distribution tests do not assume that the intervening variable effect is normally distributed, consistent with the unique distribution of the product of two random, normal variables (Craig, 1936), but they do assume that individual regression coefficients are normally distributed.

The difference in the statistical background of the Clogg et al. (1992) and Freedman and Schatzkin (1992) tests and the distribution of products tests, $P = z_\alpha z_\beta$ and $z' = \alpha\beta/\sigma_{\alpha\beta}$, makes the similarity of the empirical power and Type I error rates when α = β = 0 somewhat surprising.

The Clogg et al. (1992) and Freedman and Schatzkin (1992) tests underestimate the standard errors, which serves to compensate for critical values that are too low when standard reference distributions are used. Although the degree of compensation appeared to be quite good under some of the conditions investigated in the present simulation, it is unclear whether these tests could be expected to show an appropriate degree of compensation under other conditions (e.g., larger effect sizes and other levels of significance).

There is an important exception to the accuracy of the Clogg et al. (1992), Freedman and Schatzkin (1992), $P = z_\alpha z_\beta$, and $z' = \alpha\beta/\sigma_{\alpha\beta}$ tests. When the true population values are $\alpha = 0$ and $\beta \neq 0$, the methods lead to the conclusion that there is an intervening variable effect far too often, although the distribution of products test $z' = \alpha\beta/\sigma_{\alpha\beta}$ is less susceptible to Type I errors than the other methods. When the true value of $\alpha \neq 0$ and $\beta = 0$, the Clogg et al. (1992) and Freedman and Schatzkin (1992) tests still perform well and the two distribution methods give Type I errors that are too high. The better performance for the Clogg et al. (1992) test when $\alpha \neq 0$ and $\beta = 0$ is not surprising because the test of significance is equivalent to the test of whether the β parameter is statistically significant (H_0: $\beta = 0$) and does not include the α value in the test. The test based on the empirical distribution of $z' = \alpha\beta/\sigma_{\alpha\beta}$ has the lowest Type I error rates of the four best methods when looking across both the $\alpha = 0$ and $\beta \neq 0$ and the $\alpha \neq 0$ and $\beta = 0$ cases.

In summary, statistical tests of the intervening variable effect trade off two competing problems. First, the nonnormal sampling distribution of the $\alpha\beta$ effect leads to tests that are associated with empirical levels of significance that are lower than the stated levels when H_0 is true as well as low statistical power when H_0 is false. The MacKinnon et al. (1998) z' and P tests explicitly address this problem and provide accurate Type I error rates when $\alpha = \beta = 0$ and relatively high levels of statistical power when H_0 is false. Second, the test of the null hypothesis for $\alpha\beta = 0$ is complex because the null hypothesis takes a compound form, encompassing (a) $\alpha = 0$, $\beta = 0$; (b) $\alpha \neq 0$, $\beta = 0$; and (c) $\alpha = 0$, $\beta \neq 0$. Two of the MacKinnon et al. tests break down and yield higher than stated Type I error rates under conditions b and c. In contrast, the use of otherwise inappropriate conservative critical values based on the normal sampling distribution turns out empirically to compensate for the inflation in the Type I error rate associated with the compound form of the null hypothesis.

Statistical Recommendations

Focusing initially on the statistical performance of the tests of the intervening variable effect, the 14 tests can be divided into three groups of tests with similar performance. The first group consists of the Baron and Kenny (1986) and Judd and Kenny (1981b) approaches, which have low Type I error rates and the lowest statistical power in all conditions studied. Four tests are included in the second group of methods consisting of $P = z_\alpha z_\beta$ and $z' = \alpha\beta/\sigma_{\alpha\beta}$, the Clogg et al. (1992), and Freedman and Schatzkin (1992) tests, which have the greatest power when both α and β are nonzero and the most accurate Type I error rates when both α and β are zero. These four methods can be ordered from the best to the worst as $z' = \alpha\beta/\sigma_{\alpha\beta}$, $P = z_\alpha z_\beta$, Freedman and Schatzkin test, and Clogg et al. test for most values of α and β. If researchers wish to have the maximum power to detect the intervening variable effect and can tolerate the increased Type I error rate if *either* the α or β population parameter is zero, then these are the methods of choice. If there is evidence that $\alpha \neq 0$ and $\beta = 0$, then the Clogg et al. and Freedman and Schatzkin methods will have increased power and accurate Type I error rates and $P = z_\alpha z_\beta$ and $z' = \alpha\beta/\sigma_{\alpha\beta}$ tests have Type I error rates that are too high. When $\alpha \neq 0$ and $\beta = 0$, the Clogg et al., and the Freedman and Schatzkin methods have very high Type I error rates. For both cases where either α or β is zero, the empirical distribution method $z' = \alpha\beta/\sigma_{\alpha\beta}$ has the lowest Type I error rates (Type I error rates did not exceed .426). As a result, if the researcher seeks the greatest power to detect an effect and does not consider an effect to be transmitted through an

intervening variable if α can be zero, then the z' = $\alpha\beta/\sigma_{\alpha\beta}$ empirical distribution test is the test of choice. The researcher should be aware that the Type I error rate can be higher than nominal values for the situation whether either α or β (but not both) is zero in the population.

Eight tests are included in the third group of methods, which represent less power and too low Type I error rates when $\alpha = \beta = 0$, but more accurate Type I error rates when either α or β is zero. The tests listed in terms of accuracy consist of the joint significance test of α and β, asymmetric critical value test, test of the simple minus partial correlation, test of product of correlations, unbiased test of αβ, first-order tests of αβ, McGuigan and Langholtz (1988) test of $\tau - \tau'$, and then second-order test of αβ. Unfortunately, Goodman's (1960) unbiased test often yields negative variances and is hence undefined for zero or small effects or small sample sizes. The joint significance test of α and β appears to be the best test in this group as it has the most power and the most accurate Type I error rates in all cases compared to the other methods. Note that no parameter estimate or standard error of the intervening variable effect is available for the joint test of the significance of α and β so that effect sizes and confidence intervals are not directly available. Consequently, other tests that are close to the joint significance test in accuracy such as the asymmetric confidence interval test may be preferable as they do include an estimate of the magnitude of the intervening variable effect. The very close simulation performance of the other six methods in this group suggests that for practical data analysis, the choice of tests in this group will not change the conclusions of the study. Overall, the methods in the third group represent a compromise with less power than some methods, and more accurate Type I error rates than other methods.

The overall pattern of results for the case of $\alpha = 0$ and $\beta \neq 0$ forces consideration of two different statistical null hypotheses regarding intervening variable effects. The first hypothesis is a test of whether the indirect effect αβ is zero. This hypothesis is best tested by the methods with the most power, empirical distribution z' = $\alpha\beta/\sigma_{\alpha\beta}$ and distribution of $P = z_\alpha z_\beta$, and possibly by the Freedman and Schatzkin (1992) test. The second hypothesis is a test of whether both paths α and β are equal to zero. In this case the joint significance of α and β or the asymmetric confidence limit test provide the most direct test of the hypothesis. Given the importance of establishing that (a) the treatment leads to changes in the intervening variable ($\alpha \neq 0$) and (b) the intervening variable is associated with dependent variable ($\beta \neq 0$) (see Krantz, 1999), we strongly recommend this test for experimental investigations involving the simple intervening variable model portrayed in Figure 1. Asymmetric confidence limits for the mediated effect, αβ, can then be computed based on the distribution of the product.

Causal Inference

This article has focused on the statistical properties of intervening effect tests, at least in part because the requirements for causal inference regarding an intervening effect are complex and controversial. The majority of the statistical tests of intervening effects reported in psychology journals test the significance of the indirect effect αβ or each of the causal steps proposed by Judd and Kenny (1981b) or Baron and Kenny (1986). The tests of the indirect effect largely follow the tradition of path analysis in which a restricted model is hypothesized, here $X \rightarrow I \rightarrow Y$ and the hypothesized model is tested against the data. Although important competing models may be considered and also tested against the data, this tradition typically seeks to demonstrate only that the causal processes specified by the hypothesized model are consistent with the data.

In contrast, Judd and Kenny (1981b) originally presented their causal steps method as a device for directly probing the causal process through which a treatment produces an outcome. The strength of their approach is that in the context of a single randomized experiment it provides evidence that the treatment causes the intervening variable, the treatment causes the outcome, and that the data are consistent with the proposed intervening variable model $X \rightarrow I \rightarrow Y$, where

X represents the treatment conditions. However, the third causal step makes the strong assumption that the residuals ε_2 and ε_3 in Equations 2 and 3, respectively, are independent. This assumption can be violated for several reasons including omitting a variable from the path model, an interaction between *I* and *X*, incorrect specification of the functional form of the relations, error of measurement in the intervening variable, and a bidirectional causal relation between *I* and *Y* (Baron & Kenny, 1986; MacKinnon, 1994). When this assumption is violated, biased estimates of the indirect effect may be obtained and the causal inference that $X \rightarrow I \rightarrow Y$ may be unwarranted. Holland (1988) presents an extensive analysis of the assumptions necessary for causal inference in this design. Among the conditions necessary for causal inference are randomization, linear effects, and that the full effect of the treatment operates through the intervening variable (i.e., no partial intervening variable effect).

Establishing the conditions necessary for causal inference requires a more complex design than the two group randomized experiment considered by Judd and Kenny (1981b). Designs in which both the treatment and intervening variable are manipulated in a randomized experiment can achieve stronger causal inferences. For example, imagine a hypothesized model in which commitment leads to intentions, which, in turn, leads to behavior. Subjects could be randomly assigned to a high or low commitment to exercise program condition, following which their intentions to exercise would be measured. Following this, subjects could be randomly assigned to a condition in which the same exercise program was easy versus difficult to access and the extent of their behavioral compliance with the program could be measured. Adding design features like randomization and temporal precedence can powerfully rule out alternative causal explanations (Judd & Kenny, 1981b; Shadish, Cook, & Campbell, 2002; West & Aiken, 1997; West, Biesanz, & Pitts, 2000).

The addition of design features can change the assumptions that are necessary for causal inference. For example, Judd and Kenny (1981b) and Baron and Kenny (1986) ruled out models in which the treatment could not be shown to affect the outcome in Equation 1. This condition rules out inconsistent effect models in which the intervening variable effect ($\alpha\beta$) and the direct effect (τ') in Figure 1 have opposite signs and may cancel out. However, if the strength of the direct path and each link of the indirect path can be manipulated in a randomized experiment, strong causal inferences can potentially be reached. A recent experiment by Sheets and Braver (1999) presented an illustration of this approach.

Conclusion

Tests of the intervening variable effect are useful because they examine processes by which variables are related. In clinical and community research, such tests are critical for the elucidation of how prevention and treatment programs work. In experimental research, such tests are critical for establishing the plausibility of causal sequences implied by theory. Reflecting the lack of consensus in the definition of an intervening effect across disciplines, the available tests address several different null hypotheses. The available procedures also differ in the extent to which they simply test whether the data are consistent with a hypothesized intervening variable model versus attempt to establish other logical features that support causal inference by ruling out other competing models. As a result of these differing conceptual bases, different assumptions, and different estimation methods, the available tests show large differences in their Type I error rates and statistical power. The present article provides researchers with more information about both the conceptual bases and the statistical performance of the available procedures for determining the statistical significance of an intervening variable effect. Our hope is that researchers will now have guidance in selecting a test of the intervening variable effect that addresses their question of interest with the maximal statistical performance.

Acknowledgments

This research was supported by U.S. Public Health Service Grant DA09757 to David P. MacKinnon. We thank Sandy Braver for comments on this research.

Appendix A

The Multivariate Delta Method

Two tests of intervening variable effects use a standard error derived using the multivariate delta method. The multivariate delta method solution for the standard error is obtained by pre- and postmultiplying the vector of partial derivatives of a function by the covariance matrix for the correlations in the function (Olkin & Finn, 1995; Sobel, 1982).

The multivariate delta method assumes a function $u = f(\nu_1,\nu_2,\nu_3)$, where (ν_1, ν_2, ν_3) has covariance matrix

$$\Sigma_\nu = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}.$$

Let (p_1, p_2, p_3) denote the partial derivatives $(\partial u/\partial \nu_1, \partial u/\partial \nu_2, \partial u/\partial \nu_3)$ of u with respect to (ν_1, ν_2, ν_3). According to the Delta method, the variance of u can be approximated by

$$Var(u) \approx \sum_{i=1}^{3}\sum_{j=1}^{3} p_i \sigma_{ij} p_j.$$

The standard error is then taken as the square root of Var(u).

Olkin and Finn (1995) derived the asymptotic covariance matrix (see Olkin & Siotani, 1976, for asymptotic results) of $(\rho_{XI}, \rho_{XY}, \rho_{IY})$. The variances and covariances among the elements of this correlation matrix (Olkin & Finn, 1995) are

	ρ_{XI}	ρ_{XY}	ρ_{IY}
ρ_{XI}	$\mathrm{var}(\rho_{XI})$		
ρ_{XY}	$\mathrm{cov}(\rho_{XI}, \rho_{XY})$	$\mathrm{var}(\rho_{XY})$	
ρ_{IY}	$\mathrm{cov}(\rho_{XI}, \rho_{IY})$	$\mathrm{cov}(\rho_{IY}, \rho_{XY})$	$\mathrm{var}(\rho_{IY})$

The formulas to calculate the variances and covariances among the correlations based on asymptotic theory from Olkin and Siotani (1976) are

$$\mathrm{var}(\rho_{XI}) = \frac{\left(1-\rho_{XI}^2\right)^2}{N}, \tag{A1}$$

$$\mathrm{var}(\rho_{XY}) = \frac{\left(1-\rho_{XY}^2\right)^2}{N}, \tag{A2}$$

$$\mathrm{var}\left(\rho_{IY}\right)=\frac{\left(1-\rho_{IY}^{2}\right)^{2}}{N}, \tag{A3}$$

$$\mathrm{cov}\left(\rho_{XI},\rho_{XY}\right)=\frac{\frac{1}{2}\left(2\rho_{IY}-\rho_{XI}\rho_{XY}\right)\left(1-\rho_{IY}^{2}-\rho_{XI}^{2}-\rho_{XY}^{2}\right)+\rho_{IY}^{3}}{N}, \tag{A4}$$

$$\mathrm{cov}\left(\rho_{XI},\rho_{IY}\right)=\frac{\frac{1}{2}\left(2\rho_{XY}-\rho_{XI}\rho_{IY}\right)\left(1-\rho_{XI}^{2}-\rho_{XY}^{2}-\rho_{IY}^{2}\right)+\rho_{XY}^{3}}{N}, \tag{A5}$$

and

$$\mathrm{cov}\left(\rho_{IY},\rho_{XY}\right)=\frac{\frac{1}{2}\left(2\rho_{XI}-\rho_{XY}\rho_{IY}\right)\left(1-\rho_{XI}^{2}-\rho_{XY}^{2}-\rho_{IY}^{2}\right)+\rho_{XI}^{3}}{N}. \tag{A6}$$

Appendix B

An SAS Program to Calculate Standard Errors Using the Multivariate Delta Method

```
data a; input rxi rxy riy nobs;*note r corresponds to correlation which is
indicated by the Greek letter rho in the article;*x, i, and y represent the
independent, intervening, and dependent variables, respectively;*variance of
the correlations from Appendix A;vrxy = ((1-rxy*rxy)*(1-rxy*rxy))/nobs;vriy =
((1-riy*riy)*(1-riy*riy))/nobs;vrxi = ((1-rxi*rxi)*(1-rxi*rxi))/
nobs;*covariances among the correlations from Appendix A;crxyriy = (.5*(2*rxi-
rxy*riy)*(1-rxi*rxi-rxy*rxy-riy*riy)+rxi*rxi*rxi)/nobs;crxyrxi = (.5*(2*riy-
rxy*rxi)*(1-rxy*rxy-riy*riy-rxi*rxi)+riy*riy*riy)/nobs;criyrxi = (.5*(2*rxy-
riy*rxi)*(1-rxi*rxi-rxy*rxy-riy*riy)+rxy*rxy*rxy)/nobs;*olkin and
finn;*partial correlation or correlation with intervening variable
removed;rxyi = (rxy-riy*rxi)/sqrt((1-riy*riy)*(1-rxi*rxi));*difference
between simple and partial correlations from Equation 7;diff = rxy-
rxyi;*partial derivatives from Equation 8;opd1 = 1-(1/(sqrt(1-riy*riy)*sqrt(1-
rxi*rxi)));opd2 = c(rxi-rxy*riy)/((sqrt(1-rxi*rxi))*(1-riy*riy)**(1.5));opd3
= (riy-rxi*rxy)/((sqrt(1-riy*riy))*(1-rxi*rxi)**(1.5));ovar = opd1*opd1*vrxy
+opd2*opd1*crxyriy+opd3*opd1*rxyrxi+opd1*opd2*crxyriy+opd2*opd2*vriy
+opd2*opd3*criyrxi+opd1*opd3*crxyrxi+opd2*opd3*criyrxi+opd3*opd3*vrxi;ose =
sqrt(ovar);zolkin = diff/ose;polkin = 1-probnorm(zolkin);*bobko & rieck from
Equation 12;corr = rxi*(riy-rxy*rxi)/(1-rxi**2);*partial derivatives from
Equation 13;bpd1 = ((rxi*rxi*riy+riy-2*rxi*rxy)/(1-rxi*rxi)**2);bpd2 = (-
(rxi*rxi)/(1-rxi*rxi));bpd3 = (rxi/(1-rxi*rxi));bobkovar = ((bpd1**2)*vrxi)+
((bpd2**2)*vrxy)+((bpd3**2)*vriy)+(2*bpd1*bpd2*crxyrxi)+(2*bpd2*bpd3*crxyriy)
+(2*bpd1*bpd3*criyrxi);bobkose = sqrt(bobkovar);zbobko = corr/bobkose;pbobko
= 1–probnorm(abobko);cards;.14 .14 0 200;proc print;var diff ose zolkin polkin
corr bobkose zbobko pbobko;run;
```

References

Ajzen, I.; Fishbein, M. Understanding attitudes and predicting social behavior. Prentice Hall; Englewood Cliffs, NJ: 1980.

Allison PD. The impact of random predictors on comparison of coefficients between models: Comment on Clogg, Petkova, and Haritou. American Journal of Sociology 1995;100:1294–1305.

Alwin DF, Hauser RM. The decomposition of effects in path analysis. American Sociological Review 1975;40:37–47.

Aroian LA. The probability function of the product of two normally distributed variables. Annals of Mathematical Statistics 1944;18:265–271.

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

Bentler, P. EQS for Windows (Version 5.6) [Computer software]. Multivariate Software; Encino, CA: 1997.

Bobko P, Rieck A. Large sample estimators for standard errors of functions of correlation coefficients. Applied Psychological Measurement 1980;4:385–398.

Bollen, KA. Total direct and indirect effects in structural equation models.. In: Clogg, CC., editor. Sociological methodology. American Sociological Association; Washington, DC: 1987. p. 37-69.

Clogg CC, Petkova E, Cheng T. Reply to Allison: More on comparing regression coefficients. American Journal of Sociology 1995;100:1305–1312.

Clogg CC, Petkova E, Shihadeh ES. Statistical methods for analyzing collapsibility in regression models. Journal of Educational Statistics 1992;17(1):51–74.

Cohen, J. Statistical power for the behavioral sciences. Erlbaum; Hillsdale, NJ: 1988.

Cohen, J.; Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Erlbaum; Hillsdale, NJ: 1983.

Craig CC. On the frequency function of xy. Annals of Mathematical Statistics 1936;7:1–15.

Fox J. Effect analysis in structural equation models. Sociological Methods and Research 1980;9:3–28.

Freedman LS, Schatzkin A. Sample size for studying intermediate endpoints within intervention trials of observational studies. American Journal of Epidemiology 1992;136:1148–1159. [PubMed: 1462974]

Goodman LA. On the exact variance of products. Journal of the American Statistical Association 1960;55:708–713.

Hansen WB. School-based substance abuse prevention: A review of the state of the art in curriculum, 1980–1990. Health Education Research: Theory and Practice 1992;7:403–430.

Holland, PW. Causal inference, path analysis, and recursive structural equation models (with discussion).. In: Clogg, C., editor. Sociological methodology 1988. American Sociological Association; Washington, DC: 1988. p. 449-484.

James LR, Brett JM. Mediators, moderators and tests for mediation. Journal of Applied Psychology 1984;69:307–321.

Jöreskog, KG.; Sörbom, D. LISREL (Version 8.12) [Computer software]. Scientific Software International; Chicago: 1993.

Judd, CM.; Kenny, DA. Estimating the effects of social interventions. Cambridge University Press; Cambridge, England: 1981a.

Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. Evaluation Review 1981b;5:602–619.

Kenny, DA.; Kashy, DA.; Bolger, N. Data analysis in social psychology.. In: Gilbert, DT.; Fiske, ST.; Lindzey, G., editors. The handbook of social psychology. McGraw-Hill; Boston: 1998. p. 233-265.

Krantz DH. The null hypothesis testing controversy in psychology. Journal of the American Statistical Association 1999;94:1372–1381.

MacCorquodale K, Meehl PE. On a distinction between hypothetical constructs and intervening variables. Psychological Review 1948;55:95–107. [PubMed: 18910284]

MacKinnon, DP. Analysis of mediating variables in prevention and intervention research.. In: Cazares, A.; Beatty, LA., editors. Scientific methods in prevention research. U.S. Government Printing Office; Washington, DC: 1994. p. 127-153.NIDA Research Monograph 139DHHS Publication No. 94-3631

MacKinnon, DP. Contrasts in multiple mediator models.. In: Rose, J.; Chassin, L.; Presson, CC.; Sherman, SJ., editors. Multivariate applications in substance use research. Erlbaum; Mahwah, NJ: 2000. p. 141-160.

MacKinnon DP, Dwyer JH. Estimating mediated effects in prevention studies. Evaluation Review 1993;17:144–158.

MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding, and suppression effect. Prevention Science 2000;1:173–181. [PubMed: 11523746]

MacKinnon, DP.; Lockwood, C. Distribution of products tests for the mediated effect. 2001. Unpublished manuscript

MacKinnon, DP.; Lockwood, C.; Hoffman, J. A new method to test for mediation.. Paper presented at the annual meeting of the Society for Prevention Research; Park City, UT. Jun. 1998

MacKinnon DP, Warsi G, Dwyer JH. A simulation study of mediated effect measures. Multivariate Behavioral Research 1995;30:41–62.

McGuigan, K.; Langholtz, B. A note on testing mediation paths using ordinary least-squares regression. 1988. Unpublished note

Meeker, WQ.; Cornwell, LW.; Aroian, LA. Selected tables in mathematical statistics, Vol. VII: The product of two normally distributed random variables. American Mathematical Society; Providence, RI: 1981.

Olkin I, Finn JD. Correlation redux. Psychological Bulletin 1995;118:155–164.

Olkin, I.; Siotani, M. Asymptotic distribution of functions of a correlation matrix.. In: Ikeda, S., editor. Essays in probability and statistics. Shinko Tsusho; Tokyo: 1976. p. 235-251.

Sampson CB, Breunig HL. Some statistical aspects of pharmaceutical content uniformity. Journal of Quality Technology 1971;3:170–178.

Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental designs for generalized causal inference. Houghton-Mifflin; Boston: 2002.

Sheets VL, Braver SL. Organizational status and perceived sexual harassment: Detecting the mediators of a null effect. Personality and Social Psychology Bulletin 1999;25:1159–1171.

Sobel, ME. Asymptotic confidence intervals for indirect effects in structural equation models.. In: Leinhardt, S., editor. Sociological methodology 1982. American Sociological Association; Washington, DC: 1982. p. 290-312.

Sobel, ME. Direct and indirect effects in linear structural equation models.. In: Long, JS., editor. Common problems/proper solutions. Sage; Beverly Hills, CA: 1988. p. 46-64.

Springer MD, Thompson WE. The distribution of independent random variables. SIAM Journal on Applied Mathematics 1966;14:511–526.

Stacy AW, Leigh BC, Weingardt KR. Memory accessibility and association of alcohol use and its positive outcomes. Experimental and Clinical Psychopharmacology 1994;2:269–282.

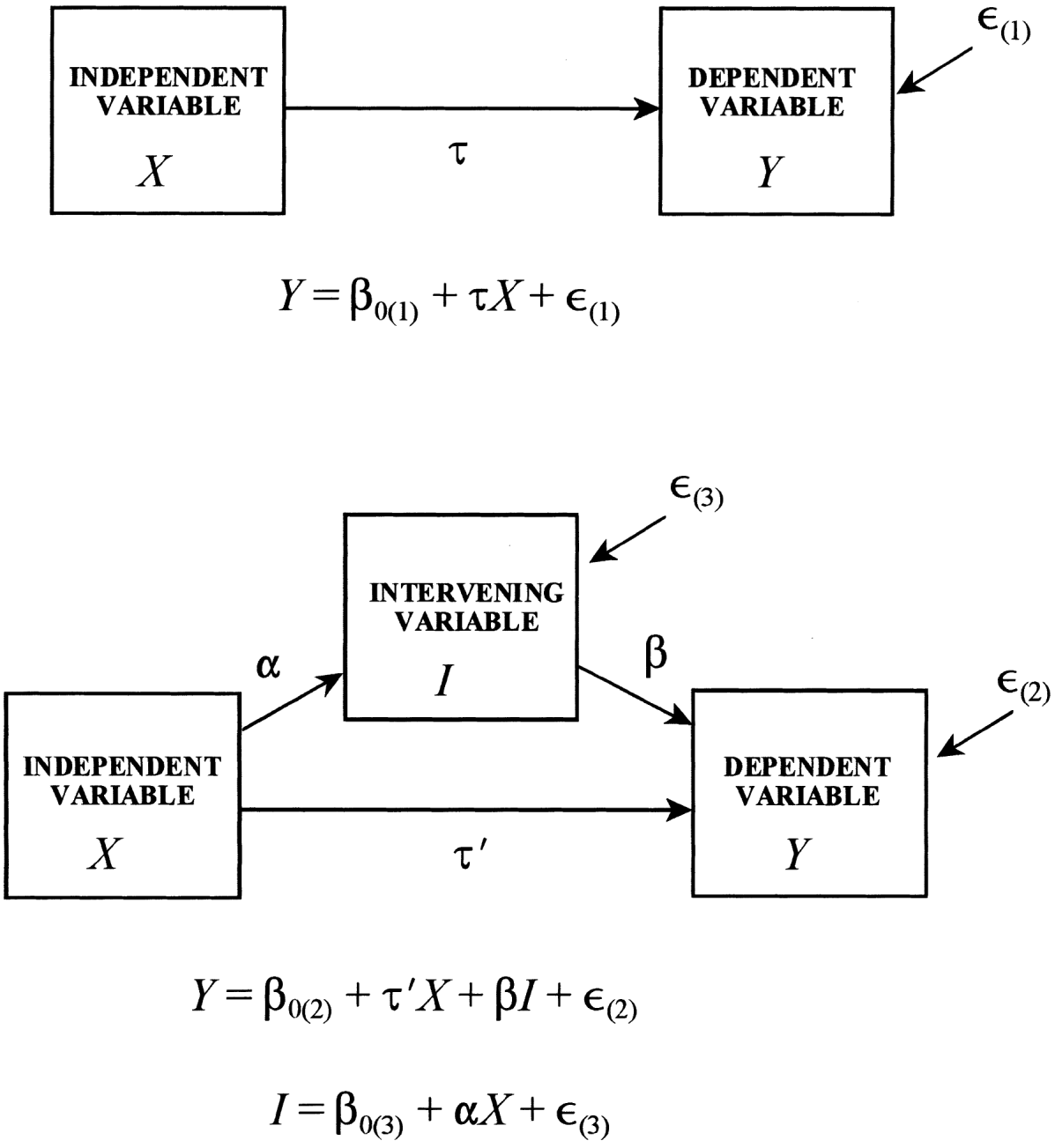
Stone CA, Sobel ME. The robustness of estimates of total indirect effects in covariance structure models estimated by maximum likelihood. Psychometrika 1990;55:337–352.

West, SG.; Aiken, LS. Toward understanding individual effects in multicomponent prevention programs: Design and analysis strategies.. In: Bryant, KJ.; Windle, M.; West, SG., editors. The science of prevention: Methodological advances from alcohol and substance abuse research. American Psychological Association; Washington, DC: 1997. p. 167-209.

West, SG.; Biesanz, JC.; Pitts, SC. Causal inference and generalization in field settings: Experimental and quasi-experimental designs.. In: Reis, HT.; Judd, CM., editors. Handbook of research methods in social and personality psychology. Cambridge University Press; New York: 2000. p. 40-84.

Woodworth, RS. Dynamic psychology.. In: Murchison, C., editor. Psychologies of 1925. Clark University Press; Worcester, MA: 1928. p. 111-126.

Yang, MCK.; Robertson, DH. Understanding and learning statistics by computer. World Scientific; Singapore: 1986.



$$Y = \beta_{0(1)} + \tau X + \epsilon_{(1)}$$

$$Y = \beta_{0(2)} + \tau' X + \beta I + \epsilon_{(2)}$$

$$I = \beta_{0(3)} + \alpha X + \epsilon_{(3)}$$

Figure 1.
Path diagram and equations for the intervening variable model.

Table 1

Summary of Tests of Significance of Intervening Variable Effects

Type of method	Estimate	Test of significance
		Causal steps
Judd & Kenny (1981a, 1981b)	None	$t_{N-2} = \frac{\tau}{\sigma_\tau}$, $t_{N-3} = \frac{\beta}{\sigma_\beta}$, $t_{N-2} = \frac{a}{\sigma_a}$, $\tau' = 0$
Baron & Kenny (1986)	None	$t_{N-2} = \frac{\tau}{\sigma_\tau}$, $t_{N-3} = \frac{\beta}{\sigma_\beta}$, $t_{N-2} = \frac{a}{\sigma_a}$
Joint significance of α and β	None	$t_{N-3} = \frac{\beta}{\sigma_\beta}$, $t_{N-2} = \frac{a}{\sigma_a}$
		Difference in coefficients
Freedman & Schatzkin (1992)	$\tau - \tau'$	$t_{N-2} = \frac{\tau - \tau'}{\sqrt{\sigma_\tau^2 + \sigma_{\tau'}^2 - 2\sigma_\tau \sigma_{\tau'}\sqrt{1 - \rho_{XI}^2}}}$
McGuigan & Langholtz (1988)	$\tau - \tau'$	$t_{N-2} = \frac{\tau - \tau'}{\sqrt{\sigma_\tau^2 + \sigma_{\tau'}^2 - 2\left(\rho_{\tau\tau'}\sigma_\tau \sigma_{\tau'}\right)}}$
Clogg et al. (1992)	$\tau - \tau'$	$t_{N-3} = \frac{\tau - \tau'}{\lvert \rho_{XI}\sigma_{\tau'} \rvert}$
Olkin & Finn (1995) simple minus partial correlation	$\rho_{XY} - \rho_{XY.I}$	$z = \frac{\rho_{XY} - \rho_{XYI}}{\sigma_{\text{Olkin \& Finn}}}$
		Product of coefficients
Sobel (1982) first-order solution	αβ	$z = \frac{a\beta}{\sqrt{a^2\sigma_\beta^2 + \beta^2\sigma_a^2}}$
Aroian (1944) second-order exact solution	αβ	$z = \frac{a\beta}{\sqrt{a^2\sigma_\beta^2 + \beta^2\sigma_a^2 + \sigma_a^2\sigma_\beta^2}}$
Goodman (1960) unbiased solution	αβ	$z = \frac{a\beta}{\sqrt{a^2\sigma_\beta^2 + \beta^2\sigma_a^2 - \sigma_a^2\sigma_\beta^2}}$
MacKinnon et al. (1998) distribution of products	$z_\alpha z_\beta$	$P = z_\alpha z_\beta$
MacKinnon et al. (1998) distribution of $\alpha\beta/\sigma_{\alpha\beta}$	αβ	$z' = \frac{a\beta}{\sqrt{a^2\sigma_\beta^2 + \beta^2\sigma_a^2}}$
MacKinnon & Lockwood (2001) asymmetric distribution of products	αβ	$a\beta \pm \text{CL}\sqrt{a^2\sigma_\beta^2 + \beta^2\sigma_a^2}$

Type of method	Estimate	Test of significance
Bobko & Rieck (1980) product of correlations	$\frac{\rho_{XI}\left(\rho_{IY}-\rho_{XY}\rho_{XI}\right)}{\left(1-\rho_{XI}^{2}\right)}$	$z=\frac{\frac{\rho_{XI}\left(\rho_{IY}-\rho_{XY}\rho_{XI}\right)}{\left(1-\rho_{XI}^{2}\right)}}{\sigma_{\text{Bobko \& Rieck}}}$

Table 2

Comparison of Estimates of the Standard Error of $\tau - \tau' = \alpha\beta$

	Sample size				
Effect size	**50**	**100**	**200**	**500**	**1,000**
			Standard deviation for $\tau - \tau' = \alpha\beta$		
Zero	.0224	.0121	.0049	.0022	.0009
Small	.0376	.0214	.0162	.0089	.0062
Medium	.0855	.0549	.0386	.0251	.0184
Large	.1236	.0857	.0585	.0366	.0257
			Freedman & Schatzkin (1992)		
Zero	.0171	.0083	.0041	.0016	.0008
Small	.0238	.0154	.0107	.0062	.0045
Medium	.0598	.0402	.0282	.0178	.0126
Large	.0903	.0622	.0444	.0274	.0193
			McGuigan & Langholtz (1988)		
Zero	.0342	.0169	.0082	.0033	.0016
Small	.0431	.0260	.0165	.0093	.0065
Medium	.0867	.0577	.0400	.0250	.0175
Large	.1252	.0861	.0603	.0374	.0264
			Clogg et al. (1992)		
Zero	.0168	.0083	.0041	.0016	.0008
Small	.0233	.0152	.0107	.0062	.0045
Medium	.0579	.0392	.0276	.0175	.0123
Large	.0859	.0595	.0428	.0264	.0186
			Sobel (1982) first order		
Zero	.0264	.0129	.0062	.0025	.0012
Small	.0371	.0236	.0156	.0090	.0064
Medium	.0841	.0568	.0397	.0249	.0175
Large	.1235	.0855	.0601	.0374	.0264

Effect size	Sample size				
	50	100	200	500	1,000
			Aroian (1944) second order		
Zero	.0348	.0170	.0082	.0033	.0016
Small	.0435	.0261	.0165	.0093	.0065
Medium	.0869	.0577	.0400	.0249	.0175
Large	.1253	.0861	.0603	.0374	.0264
			Goodman (1960) Unbiased		
Zero	.0257	.0135	.0060	.0026	.0012
Small	.0368	.0224	.0148	.0088	.0063
Medium	.0814	.0558	.0393	.0248	.0175
Large	.1217	.0848	.0599	.0373	.0264

Note. The measure of the true standard error is the standard deviation of $\tau - \tau' = \alpha\beta$. The Goodman (1960) standard error was undefined (negative variance) for the zero effect, 185, 203, 195, 208, and 190 times for samples sizes of 50, 100, 200, 500, and 1,000, respectively. The Goodman standard error was undefined for the small effect size 106, 38, and 5 times for sample sizes of 50, 100, and 200, respectively, and undefined 1 time for medium effect and a sample size of 50.

Psychol Methods. Author manuscript; available in PMC 2010 February 10.

Table 3

Comparison of Multivariate Delta Standard Error Estimates to Standard Deviation of Point Estimates

	Sample size				
Effect size	**50**	**100**	**200**	**500**	**1,000**
	Standard deviation for simple minus partial correlation $r_{difference}$				
Zero	.0215	.0119	.0049	.0021	.0009
Small	.0361	.0207	.0159	.0087	.0061
Medium	.0721	.0489	.0347	.0227	.0162
Large	.0947	.0623	.0456	.0288	.0197
	Olkin & Finn (1995) standard error				
Zero	.0252	.0127	.0061	.0025	.0013
Small	.0346	.0226	.0151	.0089	.0063
Medium	.0711	.0496	.0351	.0221	.0156
Large	.0921	.0649	.0461	.0270	.0205
	Standard deviation of product of coefficients for standardized variables $r_{product}$				
Zero	.0205	.0117	.0051	.0021	.0009
Small	.0349	.0207	.0157	.0088	.0061
Medium	.0698	.0465	.0323	.0215	.0157
Large	.0831	.0556	.0405	.0252	.0177
	Bobko & Rieck (1980) standard error				
Zero	.0256	.0127	.0061	.0025	.0012
Small	.0351	.0228	.0152	.0089	.0063
Medium	.0722	.0497	.0351	.0219	.0155
Large	.0919	.0637	.0452	.0282	.0199

Note. The measure of the true standard error for the simple minus partial correlation is the standard deviation of the simple minus partial correlation. The measure of the true standard error of the product of coefficients for standardized variables is the standard deviation of the product of coefficients for standardized variables.

Table 4

Type I Error Rates and Statistical Power for Causal Step Methods

	Sample size				
Effect size	**50**	**100**	**200**	**500**	**1,000**
			Judd & Kenny (1981a, 1981b)		
Zero	0	0	.0020	0	0
Small	.0040	0	.0060	.0400	.0740
Medium	.1060	.2540	.4940	.8620	.9520
Large	.4580	.7940	.9520	.9460	.9500
			Baron & Kenny (1986)		
Zero	0	0	.0020	0	0
Small	.0040	0	.0100	.0600	.1060
Medium	.1160	.2760	.5200	.8820	.9960
Large	.4700	.8220	.9880	1.000	1.000
			Joint significance of α and β		
Zero	.0040	.0060	.0020	.0020	0
Small	.0360	.0660	.2860	.7720	.9880
Medium	.5500	.9120	1.000	1.000	1.000
Large	.9300	1.000	1.000	1.000	1.000

Note. For all analyses, α = β and τ′ = 0. Small effect size = .14, medium effect size = .36, and large effect size = .51. Tests are two-tailed, $p = .05$. For each method, values in the first row for each test are estimates of the empirical Type I error rate. Values in rows 2–4 represent empirical estimates of statistical power.

Psychol Methods. Author manuscript; available in PMC 2010 February 10.

Table 5

Type I Error Rates and Statistical Power for Difference in Coefficients Methods

	Sample size				
Effect size	**50**	**100**	**200**	**500**	**1,000**
			$\tau - \tau'$ (Freedman & Schatzkin, 1992)		
Zero	.0160	.0440	.0180	.0520	.0500
Small	.1240	.2280	.5060	.8900	.9920
Medium	.7100	.9560	1.000	1.000	1.000
Large	.9560	1.000	1.000	1.000	1.000
			$\tau - \tau'$ (McGuigan & Langholtz, 1988)		
Zero	0	0	0	0	0
Small	.0060	.0060	.0920	.5260	.9740
Medium	.3380	.8540	1.000	1.000	1.000
Large	.8920	1.000	1.000	1.000	1.000
			$\tau - \tau'$ (Clogg et al., 1992)		
Zero	.0320	.0660	.0320	.0620	.0540
Small	.1780	.2840	.5100	.8920	.9920
Medium	.7320	.9560	1.000	1.000	1.000
Large	.9580	1.000	1.000	1.000	1.000
			Simple minus partial correlation (Olkin & Finn, 1995)		
Zero	.0020	0	.0020	0	0
Small	.0100	.0120	.1260	.5780	.9800
Medium	.4340	.8920	1.000	1.000	1.000
Large	.9380	1.000	1.000	1.000	1.000

Note. For all analyses, $\alpha = \beta$ and $\tau' = 0$. Small effect size = .14, medium effect size = .36, and large effect size = .51. Tests are two-tailed, $p = .05$. For each method, values in the first row for each test are estimates of the empirical Type I error rate. Values in rows 2–4 represent empirical estimates of statistical power.

Table 6

Type I Error Rates and Statistical Power for Product of Coefficients Methods

	Sample size				
Effect size	**50**	**100**	**200**	**500**	**1,000**
First-order test (Sobel, 1982)					
Zero	0	0	.0020	0	0
Small	.0060	.0100	.1220	.5620	.9760
Medium	.3600	.8620	1.000	1.000	1.000
Large	.9020	1.000	1.000	1.000	1.000
Second-order test (Aroian, 1944)					
Zero	0	0	0	0	0
Small	.0060	.0060	.0920	.5260	.9740
Medium	.3320	.8540	1.000	1.000	1.000
Large	.8920	1.000	1.000	1.000	1.000
Unbiased test (Goodman, 1960)					
Zero	.0160	.0040	.0140	.0020	.0100
Small	.0080	.0200	.1420	.6200	.9820
Medium	.3900	.8700	1.000	1.000	1.000
Large	.9120	1.000	1.000	1.000	1.000
Distribution of products test $P = z_\alpha z_\beta$ (MacKinnon et al., 1998)					
Zero	.0620	.0760	.0420	.0660	.0400
Small	.2220	.3960	.7180	.9740	1.000
Medium	.9180	.9960	1.000	1.000	1.000
Large	1.000	1.000	1.000	1.000	1.000
Distribution of $\alpha\beta/\sigma_{\alpha\beta}$ (MacKinnon et al., 1998)					
Zero	.0560	.0680	.0400	.0600	.0420
Small	.2060	.3600	.6920	.9580	.9960
Medium	.9040	.9960	1.000	1.000	1.000
Large	.9980	1.000	1.000	1.000	1.000

Effect size	Sample size				
	50	100	200	500	1,000
Asymmetric distribution of products test (MacKinnon & Lockwood, 2001)					
Zero	.0040	.0040	.0020	0	0
Small	.0300	.0620	.2740	.7600	.9880
Medium	.5540	.9200	1.000	1.000	1.000
Large	.9400	1.000	1.000	1.000	1.000
Product of coefficients for standardized variables (Bobko & Rieck, 1980)					
Zero	.0020	0	.0020	0	0
Small	.0080	.0160	.1300	.5700	.9780
Medium	.4200	.8760	1.000	1.000	1.000
Large	.9200	1.000	1.000	1.000	1.000

Note. For all analyses, $\alpha = \beta$ and $\tau' = 0$. Small effect size = .14, medium effect size = .36, and large effect size = .51. Tests are two-tailed, $p = .05$. For each method, values in the first row for each test are estimates of the empirical Type I error rate. Values in rows 2–4 represent empirical estimates of statistical power.

Table 7

Type I Error Rates of Mixed Effects for Causal Steps Methods

	Sample size				
α value/β value	**50**	**100**	**200**	**500**	**1,000**
			Judd & Kenny (1981a, 1981b)		
Large/zero	.0440	0	0	0	0
Medium/zero	.0020	.0020	.0020	.0020	0
Small/zero	0	0	0	0	.0020
Zero/large	.0100	.0080	.0060	.0100	.0060
Zero/medium	.0480	.0040	.0040	.0020	.0080
Zero/small	0	0	0	.0080	0
			Baron & Kenny (1986)		
Large/zero	.0040	0	0	0	0
Medium/zero	.0020	.0040	.0020	0	0
Small/zero	0	0	0	0	0
Zero/large	.0020	.0060	.0080	.0020	.0040
Zero/medium	0	.0020	.0100	.0020	.0060
Zero/small	0	0	0	.0040	.0020
			Joint significance of α and β		
Large/zero	.0400	.0420	.0500	.0480	.0380
Medium/zero	.0400	.0560	.0460	.0500	.0480
Small/zero	.0100	.0200	.0300	.0380	.0360
Zero/large	.0520	.0520	.0460	.0520	.0340
Zero/medium	.0480	.0540	.0620	.0340	.0500
Zero/small	.0060	.0160	.0280	.0400	.0420

Note. For all analyses, $\tau' = 0$. Small value = .14, medium value = .39, and large value = .59. Tests are two-tailed, $p = .05$. For each method, values in each row are empirical estimates ot the Type I error rate.

Table 8

Type I Error Rates of Mixed Effects for Difference in Coefficients Methods

	Sample size				
α value/β value	**50**	**100**	**200**	**500**	**1,000**
			$\tau - \tau'$ (Freedman & Schatzkin, 1992)		
Large/zero	.0440	.0540	.0500	.0480	.0380
Medium/zero	.0460	.0560	.0460	.0500	.0480
Small/zero	.0300	.0560	.0480	.0440	.0400
Zero/large	.5680	.5800	.6000	.5820	.5980
Zero/medium	.4720	.6560	.7100	.7020	.7080
Zero/small	.1120	.1980	.3940	.7520	.8840
			$\tau - \tau'$ (McGuigan & Langholtz, 1988)		
Large/zero	.0200	.0380	.0440	.0440	.0360
Medium/zero	.0120	.0280	.0360	.0440	.0440
Small/zero	0	.0020	.0080	.0080	.0140
Zero/large	.0300	.0380	.0400	.0480	.0320
Zero/medium	.0120	.0260	.0420	.0240	.0440
Zero/small	0	0	.0020	.0100	.0180
			$\tau - \tau'$ (Clogg et al., 1992)		
Large/zero	.0460	.0540	.0500	.0480	.0380
Medium/zero	.0460	.0560	.0460	.0500	.0480
Small/zero	.0460	.0640	.0500	.0440	.0400
Zero/large	.9800	1.000	1.000	1.000	1.000
Zero/medium	.7740	.9740	1.000	1.000	1.000
Zero/small	.2020	.2980	.4900	.8660	.9860
			Simple minus partial correlation (Olkin & Finn, 1995)		
Large/zero	.0340	.0380	.0480	.0380	.0364
Medium/zero	.0160	.0340	.0380	.0440	.0500
Small/zero	0	.0040	.0080	.0100	.0160
Zero/large	.0340	.0380	.0540	.0520	.0280

	Sample size				
α value/β value	**50**	**100**	**200**	**500**	**1,000**
Zero/medium	.0300	.0400	.0040	.0280	.0500
Zero/small	0	.0060	.0080	.0140	.0200

Note. For all analyses, $\tau' = 0$. Small value = .14, medium value = .39, and large value = .59. Tests are two-tailed, $p = .05$. For each method, values in each row are empirical estimates of the Type I error rate.

Table 9

Type I Error Rates of Mixed Effects for Product of Coefficients Methods

	Sample size				
α value/β value	**50**	**100**	**200**	**500**	**1,000**
First-order test (Sobel, 1982)					
Large/zero	.0240	.0460	.0460	.0460	.0360
Medium/zero	.0120	.0300	.0380	.0460	.0440
Small/zero	0	.0020	.0080	.0100	.0140
Zero/large	.0320	.0400	.0400	.0500	.0320
Zero/medium	.0200	.0300	.0420	.0240	.0440
Zero/small	0	.0020	.0080	.0160	.0220
Second-order test (Aroian, 1944)					
Large/zero	.0200	.0380	.0440	.0440	.0360
Medium/zero	.0120	.0280	.0360	.0440	.0440
Small/zero	0	.0020	.0080	.0080	.0140
Zero/large	.0300	.0380	.0400	.0480	.0320
Zero/medium	.0120	.0260	.0420	.0240	.0440
Zero/small	0	0	.0020	.0100	.0180
Unbiased test (Goodman, 1960)					
Large/zero	.0280	.0480	.0480	.0480	.0360
Medium/zero	.0160	.0320	.0400	.0460	.0480
Small/zero	.0220	.0080	.0100	.0120	.0140
Zero/large	.0380	.0420	.0420	.0500	.0320
Zero/medium	.0300	.0360	.0480	.0260	.0440
Zero/small	.0080	.0160	.0100	.0200	.0240
Distribution of products test $P = z_\alpha z_\beta$ (MacKinnon et al., 1998)					
Large/zero	.5860	.6720	.8080	.8820	.8860
Medium/zero	.2940	.5340	.6600	.8820	.8740
Small/zero	.1160	.1720	.2580	.4340	.5920
Zero/large	.6260	.6700	.7920	.8580	.9120

	Sample size				
α value/β value	**50**	**100**	**200**	**500**	**1,000**
Zero/medium	.4320	.5400	.6780	.8180	.8700
Zero/small	.1380	.1800	.2800	.4560	.5680
Distribution of $\alpha\beta/\sigma_{\alpha\beta}$ (MacKinnon et al, 1998)					
Large/zero	.3460	.3520	.4260	.3760	.3600
Medium/zero	.2940	.3160	.3440	.3500	.3660
Small/zero	.0900	.1480	.1920	.2800	.3100
Zero/large	.3520	.3700	.3660	.3860	.3660
Zero/medium	.3080	.3380	.3780	.3500	.3800
Zero/small	.1360	.1680	.2380	.3260	.2860
Asymmetric distribution of products test (MacKinnon & Lockwood, 2001)					
Large/zero	.0280	.0380	.0480	.0480	.0400
Medium/zero	.0240	.0420	.0400	.0460	.0440
Small/zero	.0060	.0120	.0160	.0280	.0280
Zero/large	.0440	.0360	.0400	.0460	.0340
Zero/medium	.0300	.0340	.0380	.0320	.0460
Zero/small	.0060	.0100	.0160	.0260	.0200
Product of coefficients for standardized variables (Bobko & Rieck, 1980)					
Large/zero	.0340	.0500	.0500	.0480	.0360
Medium/zero	.0200	.0320	.0400	.0460	.0480
Small/zero	0	.0020	.0080	.0100	.0140
Zero/large	.0420	.0480	.0420	.0500	.0320
Zero/medium	.0300	.0380	.0480	.0280	.0440
Zero/small	0	.0080	.0080	.0180	.0220

Note. For all analyses, $\tau' = 0$. Small value = .14, medium value = .39, and large value = .59. Tests are two-tailed $p = .05$. For each method, values in each row are empirical estimates of the Type I error rate.