Published in final edited form as:

Nurs Res. 2010 ; 59(4): 288–294. doi:10.1097/NNR.0b013e3181dd26b3.

Testing Mediation in Nursing Research: Beyond Baron and Kenny

Melanie R. Krause, MS, RN [Predoctoral Fellow],

School of Nursing, University of Wisconsin-Madison, Madison, Wisconsin

Ronald C. Serlin, PhD [Professor Emeritus],

Department of Educational Psychology, University of Wisconsin-Madison, Madison, Wisconsin

Sandra E. Ward, PhD, RN, FAAN [Helen Denne Schulte Professor of Nursing],

School of Nursing, University of Wisconsin-Madison, Madison, Wisconsin

Rachel Yaffa Zisk Rony, PhD, MPH, RN [Postdoctoral Fellow],

Department of Family Medicine, University of Wisconsin-Madison, Madison, Wisconsin

Miriam O. Ezenwa, MS, RN [Predoctoral Fellow], and

School of Nursing, University of Wisconsin-Madison, Madison, Wisconsin

Florence Naab, MS, RN [Doctoral Student]

School of Nursing, University of Wisconsin-Madison, Madison, Wisconsin

Abstract

Background—Baron and Kenny (1986) defined mediation and described how to perform statistical tests of mediation hypotheses. Their approach to testing mediation has been used extensively in the nursing literature. However, many statisticians have identified problems with the Baron and Kenny approach.

Purpose—To present alternative approaches to testing mediation.

Approach—The Baron and Kenny approach and its shortcomings are briefly reviewed. A critical analysis is then presented of 17 alternate methods in three categories: (a) causal steps, (b) difference in coefficients, and (c) product of coefficients. The analysis was focused on Type I error rate control, power, ease of computation, and versatility of use.

Results—Of the methods that control Type I error rate adequately, the joint significance test of α and β , the asymmetric distribution of products test, and the test of the products using the percentile Bootstrap method are the most powerful tests of mediation. Of these three, the joint significance test of α and β is superior due to its computational ease and versatility of use.

Discussion—Knowledge development in nursing will benefit from continued research testing mediation models. Nurse researchers could move beyond the Baron and Kenny approach to utilize more robust tests of mediation.

Correspondence concerning this article should be addressed to Sandra Ward, School of Nursing, University of Wisconsin-Madison, Box 2455, Clinical Sciences Center, K6/3, 600 Highland Avenue, Madison, Wisconsin, 53792, sward@wisc.edu. (Rachel) Yaffa Zisk Rony is now at Henrietta Szold Hadassah-Hebrew University, School of Nursing, Jerusalem, Israel and Hadassah University Hospital, Jerusalem, Israel.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

models, statistical; data analysis, statistical; mediation; data interpretation, statistical

Understanding how and why an independent variable (X) influences a dependent variable (Y) is critical in both descriptive and intervention research. In descriptive studies the concern is often with testing models that explain health and illness processes (e.g., stress and coping processes). In intervention studies the goal is to understand not only *if* an intervention is effective but also *how* or *why* it works. Variables that explain this *how* or *why* are termed mediators (Ms). Mediator variables are used to identify the essential processes that must occur for an X to have an effect on a Y (Shadish, Cook, & Campbell, 2002). In other words, M “represents the generative mechanism through which the focal independent variable [X] is able to influence the dependent variable [Y] of interest” (Baron & Kenny, 1986, p. 1,173).

In 1986, Baron and Kenny published a paper in which they defined mediation and described how to perform statistical tests of mediation hypotheses. The paper has been widely influential in the scientific literature in general, with 12,759 citations in the Science Citation Index, 17,428 citations in Google Scholar, and 9,718 citations in PsycINFO as of December 28, 2009. Two articles describing the Baron and Kenny approach for a nursing audience have been published, one in each of two leading nursing research journals (Bennett, 2000; Lindley & Walker, 1993).

However, it is important to note that statisticians have identified problems with Baron and Kenny’s approach (Bollen & Stine, 1990; Collins, Graham, & Flaherty, 1998; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Shrout & Bolger, 2002). Even Kenny (2008) has modified his recommendations recently for testing mediation. The purposes of this article are to describe briefly the original Baron and Kenny method of testing mediation, to discuss problems with this approach, and to summarize and critically evaluate alternative approaches to testing mediation with respect to adequacy of Type I error rate, power, and computational ease and versatility.

The Original Baron and Kenny Method of Testing Mediation

Baron and Kenny (1986) provided the following guidance for determining whether a third variable functions as a mediator (Figure 1):

A variable functions as a mediator when it meets the following conditions: (a) variations in levels of the independent variable [X] significantly account for variations in the presumed mediator [M] (i.e., Path *a*), (b) variations in the mediator [M] significantly account for variations in the dependent variable [Y] (i.e., Path *b*), and (c) when Paths *a* and *b* are controlled, a previously significant relation between the independent [X] and dependent [Y] variables [Path *c* in Figure 1] is no longer significant, with the strongest demonstration of mediation occurring when Path *c* [Path *c'* in Figure 1] is zero. (p. 1,176)

In this excerpt, as is true in all mediation models, X is thought to precede M in time, and M is a plausible cause of Y. Also note that step (c) implies that the relationship (the total effect) between X and Y has been tested and found to be significant. Under ideal conditions, when all variables pertinent to the relationship between X and Y are controlled, mediation would be expected to explain the relationship completely between X and the Y (path *c'* = 0). However, in actual research, Y may be explained by more than one X, and all potentially pertinent variables are not identified, much less measured and then controlled. In these

cases, mediation could not be expected to explain the relationship completely between X and Y (path $c' \neq 0$). Such a condition is termed *partial mediation*, meaning that M only partly explains the relationship between the X and Y.

Critique of the Original Baron and Kenny Approach

Despite its influence in the scientific literature, the Baron and Kenny (1986) approach to testing mediation has two important flaws. The first flaw is the requirement that a statistically significant total effect of X on Y is demonstrated before proceeding to test for mediation. The second relates to the requirement that mediation is demonstrated if a previously significant relationship between X and Y is no longer significant once M is included.

Showing a Statistically Significant Total Effect

There are two reasons mediation might be found in the absence of a statistically significant total effect of X on Y--suppression and dilution (Shrout & Bolger, 2002). Suppression occurs when the mediating effect of a competing process has the opposite sign of the mediating effect of interest and, therefore, the two effects cancel out what would otherwise have been a significant direct effect (Figure 2).

To illustrate this using an example, consider Sheets and Braver's (1999) study of organizational status (X) and perceived sexual harassment (Y). In this study, it was hypothesized that power differentials (M) would mediate the relationship between organizational status (X) and sexual harassment experiences (Y) in the workplace. The researchers found, however, that there were two opposing Ms, the perception of the harasser's power (M) and the harasser's social dominance, which was operationalized as perceived desirability as a mate (MM). More specifically, the researchers found that higher organizational status was related to increased perceptions of power (path a), which in turn increased the perception that the behavior was harassing (path b). However, it was found also that higher organizational status was related to higher perceptions of social dominance (desirability as a mate; path d), which in turn decreased the perception that the behavior was harassing (path e). These two opposing effects cancel each other out, often yielding a small total effect (path c). In the idealized scenario mentioned previously, whereby these two Ms constitute the only pertinent variables, path c would be equal to the sum of $d * e$ and $a * b$; if the products are equal in magnitude and of opposite sign, the total effect (path c) would equal zero. In other words, there are two competing mediating processes, M and MM, with opposing signs, and these two effects cancel out what might otherwise have led to a significant total effect (path $c = a * b + d * e$).

A lack of relationship between an X and Y also can occur when Y is quite distal from X, a condition termed *dilution*. The direct effect may be moderate or large in magnitude when the causal process is temporally proximal. However, the direct effect likely will become smaller (diluted) as the proposed causal chain is longer; that is, when there are additional links in the causal chain (Shrout & Bolger, 2002).

To illustrate dilution of the impact of X on Y, consider the example of a psycho-educational intervention (X) to improve cancer pain management (Y) through reducing attitudinal barriers regarding pain management (M_1 ; Ward et al., 2009). In this study, the intervention (X) changed attitudinal barriers (M_1), which in turn changed outcomes (Y) such as pain severity and quality of life, yet there was no main effect of the intervention on these outcomes. The investigators suggested that this lack of effect could have been due to the outcomes being quite distal to the intervention and the presence of unmeasured Ms in the

causal chain, such as patient-clinician interactions (M_2) and changes in medication orders (M_3 ; Figure 3).

Requiring that a Previously Significant Relationship Become Nonsignificant

The second flaw of the Baron and Kenny (1986) approach relates to the requirement that “when Paths a and b are controlled, a previously significant relation between the independent [X] and dependent [Y] variables is no longer significant” (p. 1,176). This requirement is problematic because the reduction of a statistically significant total effect to a nonsignificant partialled direct effect (path c') may be a trivial change. For example, the total effect without controlling for M may have been significant, with a p-value of .049. After controlling for M, the partialled direct effect (path c') may no longer be significant, with a p-value of .051. Whether this change in sample p-values is indicative of mediation in the population is questionable.

Alternative Approaches for Testing Mediation

Given the problems with the approach advocated by Baron and Kenny (1986), nurse researchers may find themselves questioning whether there are viable alternatives to this exceedingly popular but flawed approach. The Current Index to Statistics was searched from 1980 to 2009 for articles describing or evaluating statistical procedures designed to test for mediation. Because it is best to compare methods on the same sets of simulated conditions in well-designed, comprehensive, Monte Carlo studies, the primary focus was on MacKinnon et al. (2002) who compared 14 methods identified from a variety of disciplines that have been proposed to test mediation. MacKinnon et al. grouped the 14 methods into three categories: (a) causal steps, (b) difference in coefficients and (c) product of coefficients.

Additionally, several authors have suggested using resampling procedures in conjunction with product of coefficients methods to detect mediation (Bollen & Stine, 1990; Shrout & Bolger, 2002). Therefore, two other studies (MacKinnon, Lockwood, & Williams, 2004; Cheung & Lau, 2008) were included that examined three common resampling methods using simulated conditions similar to those in MacKinnon et al. (2002). All of the studies simulated data in conditions that included effect sizes of zero or those that are widely considered to be small, medium, or large, and a range of sample sizes was examined.

Causal Steps

The causal steps approach to testing mediation entails a specific sequence of tests of relationships among variables, all of which generally must be significant to declare that the mediational model holds. Three variations of the causal steps approach are outlined in the work of Baron and Kenny (1986), Cohen and Cohen (1983), and Judd and Kenny (1981a, 1981b). The sequence of tests described by Judd and Kenny (1981a) and Baron and Kenny differ only slightly. Unlike Judd and Kenny (1981a), who only discussed complete mediation, Baron and Kenny argued that models in which there is partial mediation are useful and acceptable. The third variant of the causal steps approach is referred to often as the joint significance test of α and β (Cohen & Cohen, 1983; α and β are the population path coefficients estimated by a and b). This method tests whether X is related to M by predicting M from X in a regression analysis, and whether M is related to Y by predicting Y from M in a regression analysis that also includes X as a predictor. If the two paths are *jointly significant*, mediation exists.

Difference in the Coefficients

The difference in coefficients approach entails comparing the magnitudes of the relationship between X and Y before and after M is included in the model. Four variants of the difference in coefficients approach are outlined in the work of Clogg, Petkova, and Shihadeh (1992); Freedman and Schatzkin (1992); McGuigan and Langholtz (1988); and Olkin and Finn (1995). These variants differ in terms of the pairs of coefficients that are compared, including regression coefficients or correlation coefficients. The procedures also test a range of null hypotheses about the mediating variables.

Product of the Coefficients

The product of coefficients approach entails testing the significance of the effect of M by computing its magnitude (the product of the path coefficients a and b), dividing this product by its standard error, and then comparing the resulting test statistic to a critical value, or using resampling methods to derive a confidence interval. Seven variants of the product of coefficients approaches that use critical values are outlined in the work by Aroian (1944); Bobko and Rieck (1980); Goodman (1960); MacKinnon and Lockwood (2001); MacKinnon, Lockwood, and Hoffman (1998); and Sobel (1982). These variants differ according to the way in which the standard error and the critical value are calculated (Shrout & Bolger, 2002; Sobel, 1982). A discussion of the different assumptions and order of derivatives in the approximations of the standard errors is beyond the scope of this paper.

Resampling methods use observed data to generate empirical sampling distributions of a statistic, such as the product of coefficients, whose theoretical distribution is unknown or from which it is difficult to derive usable results. The more commonly known resampling methods are the Quenouille-Tukey Jackknife (Quenouille, 1956; Tukey, 1958), the percentile Bootstrap (Efron, 1979), and the bias-corrected Bootstrap (Efron & Tibshirani, 1993). In the Jackknife procedure, repeated estimates of the statistic are calculated systematically from samples obtained by leaving out one observation at a time from the observed data set. From this set of estimates of the statistic, bias and an estimate of the variance of the statistic can be obtained. The Bootstrap method is similar to the Jackknife, but the set of estimates from which confidence intervals are obtained is generated by repeatedly sampling from the observed data at random and with replacement. The percentile Bootstrap method assumes that the bootstrap distribution provides an unbiased estimator; when this assumption does not hold, an adjustment is made in the bias-corrected Bootstrap method.

Critical Evaluation of Alternative Methods for Testing Mediation

MacKinnon et al. (2002) compared 14 methods they identified on a variety of criteria, including statistical power and control of Type I error rates. They accomplished this comparison by using Monte Carlo simulations of 1,000 iterations total, 500 iterations for when the independent variable is binary and 500 for when the independent variable is continuous, for sample sizes of 50, 100, 200, 500, and 1,000. Based on these simulations, MacKinnon et al. then reported the Type I error rate and power for each of the 14 tests at each of the 5 sample sizes for a continuous scenario.

Similarly, Monte Carlo studies (Cheung & Lau, 2008; MacKinnon et al., 2004) have been used to examine whether the Jackknife, percentile Bootstrap, or bias-corrected Bootstrap procedures could yield an improvement in the statistical properties of the product of coefficient methods. MacKinnon et al. (2004) generated 1,000 replications with sample sizes of 25, 50, 100, or 200. Cheung and Lau studied the use of resampling methods to detect mediation by generating 200 iterations with sample sizes of 100, 200, and 500. In Cheung and Lau (2008), MacKinnon et al. (2002), and MacKinnon et al. (2004), the population

effect sizes for nonzero paths were set equal to .14 (small), .39 (medium), and .59 (large), based on Cohen's (1992) characterization of effect sizes for regression analysis.

Interpreting Monte Carlo Results

Monte Carlo studies, which use simulated data to assess a statistical method's true Type I error rate, have a surprisingly long history. Pearson (1929) summarized the goal of a Monte Carlo study as follows: "If we know that [the true Type I error rate, ω] equals [the nominal Type I error rate, α] (or very nearly so) and this is true for a wide range of values of α , then our control of the first source of error will be as good" when the underlying assumptions of the test are violated as when the assumptions are met (p. 262). For example, the previously cited Monte Carlo experiments studied the true Type I error rate of tests of mediation or coverage probabilities of confidence intervals when it is known that the population distribution of the product of coefficients is not normal. In such cases, we know that ω does not equal α ; the key question is how "very nearly" equal they are.

Several authors have specified "nearness" criteria that procedures need to meet in order to be considered robust. For instance, Cochran (1952) wrote that at $\alpha = .05$, ω for a robust test should lie in the range .04–.06. Bradley (1978, 1980) specified three criteria of robustness--stringent, moderate, and liberal. Bradley's stringent criterion stated that for a robust test ω should lie in a range $\alpha \pm .1\alpha$, the moderate criterion required that ω should lie in a range $\alpha \pm .2\alpha$, and his liberal criterion stated that ω should lie in a range $\alpha \pm .5\alpha$. Alternatively, Serlin (2000) proposed a criterion that varied with the nominal α , being equal to Bradley's liberal criterion when α is smaller than .0001, equal to Cochran's criterion when α equals .01, and equal to $\alpha \pm .25\alpha$ when α is equal to .05.

All of the authors cited previously who examined the robustness of resampling methods applied Bradley's (1978) liberal criterion to determine robustness. This criterion is an inadequate basis upon which to call a method robust. When Bradley (1978) examined the widely held belief in the robustness of the t and F tests, he "criticized the robustness concept for lack of quantitative definition, insufficient qualification, the highly particularistic nature of the qualifying conditions, and for biased presentation" (p. 145). He chose as his liberal criterion one that all would agree delimited robustness from nonrobustness, and he showed that the t and F tests failed to meet even this extreme criterion. Based on his arguments, procedures that barely satisfy his liberal criterion would be better characterized as being, in Bradley's (1980) consideration, "not ultraliberal," rather than "satisfactorily robust" (p. 277).

Regardless of the criterion selected to distinguish robust and nonrobust methods, one must realize that the criterion applies to the true Type I error rate ω , not the empirical Type I error rate that results from a Monte Carlo study. If a confidence interval is constructed for ω , it is found that an empirical Monte Carlo result that barely meets Bradley's (1978) liberal criterion, based on the 1,000 iterations generated in MacKinnon et al. (2004), could have resulted from an ω as large as .09. When the result that barely meets the liberal criterion is based on the reported results for 500 samples generated with continuous data in MacKinnon et al. (2002), ω could be as large as .097, and when it is based on the 200 samples generated in Cheung and Lau (2008), ω could be as large as .122. Thus, when evaluating the results of a Monte Carlo study to assess the robustness of statistical methods, the sampling variability of the empirical α must be taken into account.

Adequacy of Control of Type I Error Rate

The results of the cited Monte Carlo studies were evaluated critically by sequentially eliminating tests of mediation that failed to meet the following criteria: (a) adequacy of

control of Type I error rate, (b) power, and (c) computational ease and versatility. A summary of the evaluation of the methods can be found in Table 1. The determination regarding Type I error rate control was conducted as follows. A nominal error rate of .05 was selected. Because it is known that test assumptions are never met precisely, it was specified how far above .05 the test's true Type I error rate ω would be allowed to be before declaring the test unacceptably liberal. Bradley's (1980) moderate upper limit of .06 was selected as the cutoff. Taking sampling variability of the empirical α into account, in a Monte Carlo study like that of MacKinnon et al. (2004), the empirical Type I error rate would have to exceed .06 by 1.645 standard deviations, 1.645 being the critical value for a one-tailed test, and the standard deviation being $\sqrt{p(1-p)/N} = \sqrt{.06(.94)/500} = .0106$, before one could declare via a significance test that the true value ω exceeded .06. Therefore, if a method's empirical rate in the MacKinnon et al. (2004) study was found to be greater than $.06 + 1.645 * .0106 = .0775$, that method would be considered to have failed to control Type I error rate adequately. Similar calculations for the 1,000 replications in MacKinnon et al. (2004) and for the 200 iterations in Cheung and Lau (2008) indicate that an empirical Type I error rate that exceeded .0724 or .0876, respectively, would reveal a method whose true Type I error rate exceeded the .06 cutoff. Eliminated from further consideration were those tests that did not control Type I error rate adequately, leaving 12 methods (Table 1).

Power

Next evaluated were those 12 methods based on their respective powers for effect sizes categorized by MacKinnon et al. (2002) as small (.14), medium (.39), and large (.59; Table 1). The power achieved by all methods was inadequate (power less than .55) for small effect sizes with samples of size 200 or fewer and for a medium effect size with a sample size of 50. Further, each of the methods had good or excellent power (greater than .80) for large and medium effect sizes for sample sizes of 100 or larger and for a small effect size with a sample size of 1,000. Therefore, the methods were discriminated among by comparing power for a small effect size when $n = 500$, the combination for which there was notable variability in power. Of the 12 methods that controlled Type I error rate adequately, the joint significance test of α and β , the asymmetric distribution of products test, and the test of the products using the percentile Bootstrap method were clearly the most powerful tests of mediation.

Computational Ease and Versatility

Finally, the joint significance test of α and β , the asymmetric distribution of products test, and the test of products were compared using the percentile Bootstrap method based on computational ease and versatility of use. With respect to computational ease, the joint significance test of α and β is an attractive method because it can be computed easily using any standard statistical package, such as SPSS or SAS. Conversely, the asymmetric distribution of products test and percentile Bootstrap method can be computationally burdensome. The asymmetric distribution of products test can be calculated by hand using critical value tables that can be accessed from a website. The asymmetric distribution of products test and the test of the products using the percentile Bootstrap method can be performed using computer programs available at MacKinnon's website (<http://www.public.asu.edu/~davidpm/ripl/mediate.htm#download>).

With respect to versatility of use, the joint significance test of α and β can be used with complex models involving multiple Xs, Ys, and Ms. Additionally, a nonparametric form of the test can be used if parametric assumptions are not met. Conversely, the symmetric distribution of products test and the test of products using the percentile Bootstrap method cannot be applied easily to complex models nor can they be used readily in nonparametric analyses. Therefore, all else being equal, the joint significance test of α and β should be used

due to the clear advantage in computational ease and versatility of use (Serlin, Jacobs, & Franke, 1995).

Conclusion

Knowledge development in nursing undoubtedly will benefit from continued research testing mediation models. However, investigators are encouraged to move beyond the Baron and Kenny (1986) approach and to utilize more robust tests. In this paper, a case has been made for the use of the joint significance test of α and β . This is the best test because it yields the most power and the most accurate type I error rates in all conditions studied compared to other tests, and because it is easily computed and versatile. More debate in this area, as well as analyzing data using the suggested approach to testing mediation, is heartily encouraged.

Acknowledgments

This work was supported by a John A. Hartford Foundation Building Academic Geriatric Nursing Capacity (BAGNC) Predoctoral Scholarship (AAN # 08-131) and F31NR010039 to Melanie R. Krause, R01 NR03126 and P20 NR008987 to Sandra E. Ward, T32PH10010 to Rachel Yaffa Zisk Rony, and F31NR010820 to Miriam O. Ezenwa.

References

- Aroian LA. The probability function of the product of two normally distributed variables. *Annals of Mathematical Statistics* 1944;18:265–270.
- 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(6):1173–1182. [PubMed: 3806354]
- Bennett JA. Mediator and moderator variables in nursing research: Conceptual and statistical differences. *Research in Nursing & Health* 2000;23(5):415–420. [PubMed: 11052395]
- Bobko P, Rieck A. Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement* 1980;4(3):385–398.
- Bollen KA, Stine R. Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology* 1990;20:115–140.
- Bradley JV. Robustness? *British Journal of Mathematical and Statistical Psychology* 1978;31:144–152.
- Bradley JV. Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society* 1980;16(5):333–336.
- Cheung GW, Lau RS. Testing mediation and suppression effects of latent variables: Bootstrapping with structural equation models. *Organizational Research Methods* 2008;11(2):296–325.
- Clogg CC, Petkova E, Shihadeh ES. Statistical methods for analyzing collapsibility in regression models. *Journal of Educational Statistics* 1992;17(1):51–74.
- Cochran WG. The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics* 1952;23:315–345.
- Cohen, J.; Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum; 1983.
- Cohen J. A power primer. *Psychological Bulletin* 1992;112(1):155–159. [PubMed: 19565683]
- Collins LM, Graham JJ, Flaherty BP. An alternative framework for defining mediation. *Multivariate Behavioral Research* 1998;33(2):295–312.
- Efron B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 1979;7:1–26.
- Efron, G.; Tibshirani, R. *An introduction to the bootstrap*. New York: Champan & Hall/CRC; 1993.
- Freedman LS, Schatzkin A. Sample size for studying intermediate endpoints within intervention trials or observational studies. *American Journal of Epidemiology* 1992;136(9):1148–1159. [PubMed: 1462974]

- Goodman LA. On the exact variance of products. *Journal of the American Statistical Association* 1960;55:708–713.
- Judd, CM.; Kenny, DA. Estimating the effects of social interventions. Cambridge, England: Cambridge University Press; 1981a.
- Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* 1981b;5(5):602–619.
- Kenny, DA. Mediation. 2008 [Retrieved February 4, 2009]. from <http://davidakenny.net/cm/mediate.htm>
- Lindley P, Walker S. Theoretical and methodological differentiation of moderation and mediation. *Nursing Research* 1993;42(5):276–279. [PubMed: 8415040]
- 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. 1998.
- MacKinnon DP, Lockwood CM, Hoffmann JM, West SG, Sheets V. A comparison of methods to test the significance of mediation and other intervening variable effects. *Psychological Methods* 2002;7(1):83–104. [PubMed: 11928892]
- MacKinnon DP, Lockwood CM, Williams J. Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research* 2004;39(1):99–128. [PubMed: 20157642]
- McGuigan K, Langholtz B. A note on testing mediation paths using ordinary least-squares regression. 1988 Unpublished note.
- Olkin I, Finn JD. Correlation redux. *Psychological Bulletin* 1995;118:155–164.
- Pearson ES. The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika* 1929;12:259–285.
- Quenouille M. Notes on bias in estimation. *Biometrika* 1956;43:353–360.
- Serlin RC. Testing for robustness in Monte Carlo studies. *Psychological Methods* 2000;5(2):230–240. [PubMed: 10937332]
- Serlin, RC.; Jacobs, V.; Franke, T. Testing for mediation. Paper presented at the annual meeting of the American Educational Research Association; San Francisco, CA. 1995 Apr.
- Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental design for generalized causal inference. Boston: Houghton-Mifflin; 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(9):1159–1171.
- Shrout PE, Bolger N. Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods* 2002;7(4):422–445. [PubMed: 12530702]
- Sobel, ME. Asymptotic confidence intervals for indirect effects in structural equation models. In: Leinhardt, S., editor. *Sociological methodology*. Washington, DC: American Sociological Association; 1982. p. 290-312.
- Tukey JW. Bias and confidence in non-quite large samples. *The Annals of Mathematical Statistics* 1958;29:614.
- Ward SE, Serlin RC, Donovan HS, Ameringer SW, Hughes S, Pe-Romashko K, et al. A randomized trial of a representational intervention for cancer pain: Does targeting the dyad make a difference? *Health Psychology* 2009;28(5):588–597. [PubMed: 19751085]

Figure 1.

Diagram of statistical mediation. Baron and Kenny used “c” to refer to both the total effect from X to Y and to refer to the direct effect (the path from X to Y once M is controlled). This practice can be confusing and therefore we follow Shrout and Bolger’s convention of using c to refer to the total effect and c’ to refer to the direct effect.

Figure 2.
Diagram of suppression of statistical mediation effects.

Figure 3.
Diagram of statistical dilution.

Table 1

Empirical Type I Error Rate for Each Method for Testing Mediation and Power Estimates for those Methods that had Acceptable Error Rate Control

Category	Related Methods	True Type I Error Rate < .06	Power for small effect when $N = 500$
Causal steps	Judd & Kenny (1981a, 1981b)	Yes	.04
	Baron & Kenny (1986)	Yes	.06
	Joint significance test of α and β (Cohen & Cohen, 1983)	Yes	.77
Difference in coefficients	Freedman & Schatzkin (1992)	No	--
	McGuigan & Langholtz (1988)	Yes	.53
	Clogg et al. (1992)	No	--
	Olkin & Finn (1995) simple minus partial correlation	Yes	.58
Product of coefficients	Sobel (1982) first-order solution	Yes	.56
	Aroian (1944) second-order exact solution	Yes	.53
	Goodman (1960) unbiased solution	Yes	.62
	MacKinnon et al. (1998) distribution of products	No	--
	MacKinnon et al. (1998) distribution of $\alpha\beta/\sigma_{\alpha\beta}$	No	--
	MacKinnon & Lockwood (2001) asymmetric distribution of products	Yes	.76
	Bobko & Rieck (1980) product of the correlations	Yes	.57
	Quenouille-Tukey Jackknife (Quenouille, 1956; Tukey, 1958)	Yes	.55
	Percentile Bootstrap (Efron, 1979)	Yes	.78
	Bias-corrected Bootstrap (Efron & Tibshirani, 1993)	No	--