Experimental approaches to mediation: A new guide for assessing causal pathways

Experimental Approaches to Mediation:

A New Guide for Assessing Causal Pathways

John G. Bullock, Donald P. Green, and Shang E. Ha

Yale University

August 16, 2008

Abstract. When mediators are not experimentally manipulated, estimates of mediation effects are likely to be biased. This is so even when the analyzed data are free from measurement error, reciprocal causality, and other problems that have attracted attention since the publication of Baron and Kenny (1986). We demonstrate the bias mathematically and show, in a content analysis of recent articles, that it is likely to affect many published studies of mediation. We propose two new methods of mediation analysis, both of which require manipulation of mediators. These methods are more difficult to implement than nonexperimental approaches, but they produce unbiased estimates of mediation effects, and the success that psychologists have already had manipulating mediators suggests that they are often feasible.

[*] We thank Marilynn Brewer, Geoff Cohen, Jack Dovidio, Ruth Ditlmann, Anthony Greenwald, Andrew Hayes, Lawrence James, Alan Kazdin, Kevin King, David Mackinnon, Margaret McConnell, Elizabeth Levy Paluck, Patrick Rosopa, and Peter Salovey for comments on previous drafts. Please send comments to donald.green@yale.edu.

A common criticism of experiments is that they reveal but do not explain causal relationships. Consider an experimental demonstration that anxiety affects subjective well-being. Does the effect reflect the tendency of anxious people to avoid social interaction, negative health effects of anxiety, or something else? Or consider a related problem in a nonexperimental setting: if we conclude that poverty causes children to be less educated, what accounts for the effect? Does it exist because school quality declines with parental income, because poor children are more likely to be malnourished, because poor families move often and thereby disrupt their children's education, or because of some other factor? Questions like these imply a search for *mediators*, variables that transmit the causal effects of other variables. Mediation analysis has a long history in the natural and social sciences: see Fisher's (1935) analysis-of-covariance-based studies of crop growth, Lazarsfeld's (1955) use of the "elaboration model" to study soldiers' attitudes, and Blau and Duncan's (1967) use of path analysis to study social mobility. But mediation analysis is now more common in psychology than in any other discipline. The best-known article on the subject, Baron and Kenny (1986), is already the most frequently cited article in the history of *JPSP* (Quiñones-Vidal, López-Garcia, Peñaranda-Ortega, & Tortosa-Gil, 2004) and one of the most frequently cited in clinical, developmental, and cognitive psychology (MacKinnon, Fairchild, & Fritz, 2007). Mediation analysis is now *de rigueur* for new social psychology manuscripts.

Because of its prominence, the method advanced by Baron and Kenny (1986) has attracted more than the usual amount of scrutiny. It is not suited to small samples (Hoyle & Kenny, 1999; MacKinnon, Lockwood, & Williams, 2004; Shrout & Bolger, 2002) or to time-series data (Maxwell & Cole, 2007). It depends on strong assumptions about the directions of causal effects (Kraemer, Stice, Kazdin, Offord, & Kupfer, 2001; Mathieu & Taylor, 2006). Measurement error in the mediators can lead to bias, and corrections for error can complicate analysis (Hoyle & Kenny, 1999; McDonald, 1997). These criticisms have led to modifications of the method, but they leave the heart of the method—what Kenny (2008, p. 353) calls "four famous steps" of regression analysis—largely intact. We agree with these criticisms, but we break with them by advancing a more fundamental objection: both the Baron-Kenny method and other methods of mediation analysis are prone to producing biased estimates of mediation effects. This bias exists even in the absence of measurement error, questions about the direction of causal effects, or concerns about statistical power.

A more general problem is an impression about mediation analysis that is conveyed by the procedures and the articles that describe them: an impression that mediation analysis is easy, or at least easy enough so that it can reasonably be expected from almost every empirical social psychology investigation. Kenny (2008, p. 355) hints at this concern, noting that Baron and Kenny (1986) is "much too formulaic" and that it "seems to imply that if a series of regression equations are estimated, the researcher has a definitive answer about mediation." In this article, we argue that a "series of regression equations" alone will almost never suffice to produce a mediation analysis that is credible, let alone definitive. Mediation analysis typically requires experimental manipulation of hypothesized mediators. The experimental designs that can be used to study mediation are more complicated than the designs that are widely used to demonstrate the effects of independent variables on dependent variables. And a complete accounting of the ways in which an effect is mediated is likely to require many experiments and years of research. Trustworthy mediation analysis is possible, but it is difficult.

In place of mediation analysis with unmanipulated mediators, we suggest two experimental alternatives. One requires direct manipulation of both the treatment and the mediators. This has been suggested before (e.g., Spencer, Zanna, & Fong, 2005; Stone-Romero & Rosopa, 2008, pp. 20-21; Sigall & Mills, 1998, p. 225); our contributions are to describe a specific experimental design and to show analytically why it is free from the bias to which nonexperimental methods are prone. Of course, it is often difficult to directly manipulate mediators. For these cases, we recommend an indirect experimental approach grounded in instrumental-variables estimation, a method that is widespread in other social sciences (MacKinnon, 2008). This approach has not been used to estimate mediation effects in psychology, and Shadish, Cook, and Campbell (2002,

p. 414) suggest that it will not be used until its strengths and weaknesses are better understood. One of our aims is to promote better understanding of its strengths and weaknesses.

The approaches that we recommend demand experimental creativity. As we show, they require researchers to design targeted manipulations that affect mediators without directly affecting other variables. It is challenging to craft such manipulations. But for decades, many social psychologists have met the challenge, and we highlight exemplary work throughout this article.

The article proceeds as follows. We begin by presenting the statistical structure of the Baron-Kenny method and explaining why it is susceptible to bias. The bias is problematic given the way that the method is applied in practice, as we show in a content analysis of 50 social-psychology articles published in 2007. We proceed by elaborating an experimental method of mediation analysis and explaining why it is free from the bias to which conventional methods are prone. Because this method requires direct manipulation of mediators, which often presents practical challenges, we also describe an alternative method that combines indirect manipulation of mediators with instrumental-variables estimation. The next section confronts a likely objection to our position, and the final section reviews and concludes.

Mediation Analyses with Unmanipulated Mediators Are Prone to Bias

There are many nonexperimental ways to test whether one variable mediates the effect of another. (See MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002 for an overview.) Our criticism applies to each of them with equal force, but for simplicity of exposition, we focus on the method proposed by Baron and Kenny (1986, p. 1177). It has four steps:

1. Regress the potential mediator on the independent variable. If the coefficient of the independent variable is not statistically significant, conclude that the potential mediator does not mediate the effect of the independent variable. If the coefficient is statistically significant,

2. Regress the dependent variable on the independent variable. If the coefficient of the independent variable is not statistically significant, conclude that the potential mediator does not mediate the effect of the independent variable. If the coefficient is statistically significant,

3. Regress the dependent variable on both the independent variable and the potential mediator. If the coefficient of the potential mediator is not statistically significant, conclude that the potential mediator does not mediate the effect of the independent variable. If the coefficient of the potential mediator is statistically significant,

4. Examine the coefficient of the independent variable from the previous step. "Perfect mediation holds if the independent variable has no effect when the mediator is controlled" (Baron & Kenny, 1176).

Typically, an independent variable is mediated by several variables, some of which are not measured. In these cases, Baron and Kenny counsel that "a more realistic goal may be to seek mediators that significantly decrease [the coefficient on the independent variable] rather than eliminating the relation between the independent and dependent variables. . . . a significant reduction indicates that a given mediator is indeed potent" (Baron & Kenny, 1986, p. 1176).

In most applications, this procedure will produce biased estimates of the extent of mediation. To see why, let Y be a dependent variable, X a randomly assigned treatment, and M a potential mediator of the treatment effect. Baron and Kenny propose that we estimate three models:

$$M = i_1 + aX + e_1, \tag{1}$$

$$Y = i_2 + cX + e_2, \text{ and} \tag{2}$$

$$Y = i_3 + dX + bM + e_3. \tag{3}$$

i_1, i_2, and i_3 are intercepts. e_1, e_2, and e_3 are mean-zero error terms that represent the cumulative effect of omitted variables; for example, e_1 represents the effect on M of variables other than X. The coefficients of interest are a, b, c, and d. The effect of M on Y is b. The total effect of X on Y is c. The "direct" effect of X on Y is d. The "indirect" or "mediated" effect of X on Y is ab, or, equivalently, $c - d$.[1]

The expected value of the Baron-Kenny estimator of b is

$$E\left[\hat{b}\right] = b + \frac{\operatorname{cov}(e_1, e_3)}{\operatorname{var}(e_1)}, \tag{4}$$

and the expected value of the Baron-Kenny estimator of d is

$$E\left[\hat{d}\right] = d - a \cdot \frac{\operatorname{cov}(e_1, e_3)}{\operatorname{var}(e_1)}, \tag{5}$$

where $\operatorname{cov}(e_1, e_3)$ is the covariance of e_1 and e_3, and $\operatorname{var}(e_1)$ is the variance of e_1. (All proofs are given in the online appendix.) That is, the Baron-Kenny estimator of b is equal, in expectation, to the true value of b plus an additional quantity. Similarly, the Baron-Kenny estimator of d is equal, in expectation, to the true value of d minus an additional quantity.[2] The Baron-Kenny estimators of b and d are unbiased only when these additional quantities equal zero, i.e., only when e_1 and e_3 do not covary.

In practice, e_1 and e_3 are likely to covary when the mediator M is not manipulated, and the Baron-Kenny method is thus likely to produce biased estimates of b and d. The problem is that unobserved variables that affect M may be correlated with unobserved variables that affect Y. When this happens, e_1 and e_3 covary, producing bias in the Baron-Kenny estimates.

Stepping back from mediation analysis to the more general problem of regression specification, estimators tend to be biased when one controls for variables that are affected by the treatment, as one does in Equation 3 by controlling for both X and M. This "post-treatment bias" is the subject of a well-developed literature (Rosenbaum, 1984; Gelman & Hill, 2007, pp. 188-94), but it has largely escaped the attention of those who use mediation analysis. At root, it is one instance of an even more general rule: estimators of the parameters of regression equations are likely to be unbiased only if the predictors in those equations are exogenous, i.e., independent of the errors (e.g., Freedman, 2005, pp. 86-92; Cohen, Cohen, West, & Aiken, 2003, pp. 475-76). And in most cases, the only way to ensure that *M* is independent of the error term is to randomly assign its values. By contrast, "the benefits of randomization are generally destroyed by including post-treatment variables" that have not been manipulated (Gelman & Hill, 2007, p. 192).

Many who have written about mediation have suggested that experiments make for better causal inference than observational studies. They have also suggested, albeit in passing, that mediation analyses are likely to yield accurate estimates of mediation effects only if the predictors in the estimated models are independent of the error terms (Judd & Kenny, 1981, p. 607; Kenny, Kashy, & Bolger, 1998, p. 262; MacKinnon et al., 2002, p. 100; Stone-Romero & Rosopa, 2008, p. 342). But these mentions understate the scope of the problem and the difficulty of solving it. Our reading of the literature suggest that six points warrant greater emphasis among practitioners. First, the assumption of independence between unmanipulated mediators and error terms is usually very implausible. Second, mediation analyses with unmanipulated mediators are at odds with the discipline's long history of experimentation and skepticism about making causal inferences from observational data. Third, random assignment of the treatment alone does not solve or even ameliorate the problem. Fourth, other methods of mediation analysis, including the many extensions of Baron and Kenny (1986), do not solve or even ameliorate the problem. Fifth, controlling for unmanipulated mediators is likely not just to bias estimates of mediation but to exaggerate them. Sixth, even when both the treatment and the mediator are randomly assigned, conventional methods of mediation analysis will typically lead to incorrect conclusions about mediation.

Because errors cannot be observed, the assumption of independence cannot be justified by examination of the data: it must be justified by theory. Those who study mediation sometimes mention the importance of “strong theory” to causal inference, but they are almost always silent about the type of theory that is required. To justify the assumption of independence, a theory must tell a substantive story about what the errors are and why they are independent of the variables of interest (e.g., Berk, 2004, pp. 93-97; Angrist, Imbens, & Rubin, 1996, p. 446). If social scientists have such theories, they have been quiet about them. Put in this light, the exogeneity assumption is extraordinary: as one statistician has it, the assumption is tantamount to trusting that “Nature ran the observational study just the way we would run an experiment. We don’t have to randomize. Nature did it for us” (Freedman, 2005, p. 86).

Historically, psychologists have not been so trusting. They have preferred to run their own experiments, ensuring the independence of predictors in regression equations by randomly assigning their values. Methodologically, nothing separates psychology from other social sciences as much as psychologists’ reliance on experiments. The recent explosion of mediation analyses with data from unmanipulated mediators thus represents a departure from the experimental tradition and the skepticism that informs it.

Researchers who only randomly assign their treatment do not protect themselves from biased estimates of mediation. Random assignment can ensure that X bears no systematic relationship to e_1, e_2, and e_3, but it says nothing about whether M or Y are systematically related to those variables, and thus nothing about whether $\text{cov}(e_1, e_3) = 0$. This warning is absent from Baron and Kenny (1986) and almost all subsequent work on mediation. But it appears clearly in an earlier article, which argues that what would come to be known as the Baron-Kenny method is “likely to yield biased estimates of causal parameters. . . *even when a randomized experimental research design has been used*” (Judd & Kenny, 1981, p. 607, emphasis in original). The point seems to have escaped many who study and use mediation. For example, Wood, Goodman, Beckmann, and Cook (2008, p. 283) report an impressively detailed survey of organizational psychology articles that use mediation, but while they distinguish between nonexperimental and experimental analyses, they do not distinguish between studies in which both the treatment and the mediator are manipulated and studies in which at least one of these variables is not manipulated. From the standpoint of mediation analysis, the second distinction is the important one.

By the same token, improvements on the method of Baron and Kenny (1986) cannot keep researchers from making biased estimates of mediation effects when a mediator is endogenous, i.e., statistically dependent on the error term in the model of the dependent variable. When a mediator is endogenous, bootstrapping and other small-sample methods merely produce better standard errors for biased estimates. Techniques adapted for time-series and panel data also address only second-order problems. Of course, some phenomena call for models more complex than the one in Equations 1 through 3, and these models in turn demand more complex methods of estimation. But these methods do nothing to solve the problem of endogenous mediators; to the contrary, they invariably pile more implausible assumptions—typically, assumptions of exogeneity at multiple points in the model—on top of the single implausible assumption of exogeneity that is at issue in Equation 3.

The methods of mediation analysis that have developed since Baron and Kenny (1986) can solve real problems, but it is striking that so much attention has been paid to these problems while so little has been paid to the more fundamental problem of assuming exogenous mediators. (James, 1980 and James, Mulaik, & Brett, 1982 are notable exceptions.) We suspect that endogeneity of mediators has received little attention because it is by far “the most difficult specification error to solve” (Kenny et al., 1998, p. 262). But it is also the most important: if the mediators in our models are endogenous, the methods of estimation that have been developed in the last two decades will be powerless to keep us from making biased estimates of mediation effects.

When mediators are endogenous, estimates of mediation produced by the Baron-Kenny method will often be not just biased but inflated. To see why, recall that the bias term in the Baron-Kenny estimator of b is $\frac{\text{cov}(e_1,e_3)}{\text{var}(e_1)}$. If this term shares the sign of b—if it is positive where b is positive, or negative where b is negative—the Baron-Kenny estimates of b will be inflated. In practice, this quantity *will* usually share the sign of b. Most of the time, the factors other than X that affect M also affect Y, and affect Y in the the same direction. For example, the effect of parents' income (X) on children's income (Y) may be mediated by children's education (M): wealthier parents purchase better schooling for their children, which in turn increases children's income. Many omitted variables are likely to influence children's income and education in the *same* direction: proximity to good schools, other neighborhood characteristics, parents' and children's attitudes toward education, returns to education in the job market, and government education policies all fit this description. Omitting any of these variables from a mediation analysis will tend to bias the analysis in favor of finding mediation effects, and the more numerous the omitted variables, the greater the bias is likely to be. The same is true of most psychological mediation analyses: it is easy to think of variables that are likely to affect mediators and dependent variables in the same way, nearly impossible to measure and control for all of them. And whenever we do not control for all such variables, the Baron-Kenny estimates of b will be inflated. Estimates of the mediation effect ab will therefore be inflated, too, making mediation effects seem stronger than they are.

The preceding discussion may seem to suggest that the Baron-Kenny method and related methods will produce unbiased estimates of mediation effects if the treatment and the mediator are randomized. Unfortunately, they will not. To see why, recall that estimation of the parameters b and d is the third step in the four-step procedure detailed by Baron and Kenny. We have established that when the treatment and the mediator are randomized, estimation of these parameters is unbiased. But in this case, the first step of the Baron-Kenny procedure will lead us to conclude that M does not mediate X, because a regression of M on X will reveal that X does not affect M. (See page 4.) Other methods will produce the same conclusion, either because they include the first step from the Baron-Kenny procedure or because they permit a conclusion that M mediates X only if c, the overall effect of X on Y, differs from d, the direct effect of X on Y. (When both M and X are randomized, c will be approximately equal to d.) Thus, the only obvious case in which the Baron-Kenny method and related methods will produce unbiased estimates of b and d is a case in which conventional methods will lead us to conclude that M does not mediate the effect of X, even if it ordinarily does. To accurately estimate mediation effects, it is typically necessary to experimentally manipulate both X and M, but it is not sufficient to apply the Baron-Kenny method or related methods to such “fully experimental” data.

To see how mediation analysis with nonexperimental data can lead us astray in practice, consider Table 2, which is provided by Rubin (2004, p. 328). Four “potential outcomes” (e.g., Frangakis & Rubin, 2002) are defined: $Y(1)$ is the set of values that the dependent variable would have if all subjects were treated (that is, if $X = 1$ for all subjects); $Y(0)$ is the set of values that it would have if no subjects were treated ($X = 0$ for all subjects); and $M(1)$ and $M(0)$ are the corresponding sets of values for the mediator. In any actual study, we would observe only one set of outcomes for each subject: either $M(1)$ and $Y(1)$ or $M(0)$ and $Y(0)$, but never both. The last three columns of the table show what we would observe if subjects were randomly assigned to either the treatment or the control group with equal probability.

Comparison of the $Y(1)$ and $Y(0)$ columns in Table 2 reveals that the true effect of the treatment on Y is +1 for every subject in the sample: in every row, $Y(1) - Y(0) = 1$. This effect is completely unmediated: it is +1 regardless of the observed value of M. In other words, the direct effect of the treatment (d) is +1 and the effect of the mediator (b) is 0.

But note what happens when we apply the Baron-Kenny procedure to the observed data in Table 2. We estimate

$$M_{obs} = \hat{i}_i + \hat{a}X = 2.67 + .67X, \tag{1*}$$

$$Y_{obs} = \hat{i}_2 + \hat{c}X = 12 + 1X, \text{ and} \tag{2*}$$

$$Y_{obs} = \hat{i}_3 + \hat{d}X + \hat{b}M_{obs} = 4 - 1X + 3M_{obs}. \tag{3*}$$

The estimated effect of the treatment on the mediator is accurate: $\hat{a} = a = .67$. The estimate of the overall treatment effect on the dependent variable is also accurate: $\hat{c} = c = 1$. But the estimate of the effect of M on Y is inflated: there is no effect ($b = 0$), but the Baron-Kenny procedure indicates an effect $\left(\hat{b} = 3\right)$. The Baron-Kenny procedure thus wrongly indicates that the treatment effect is mediated by M: $\hat{a}\hat{b} = .67 \times 3 \approx 2$. Worse, the Baron-Kenny estimate of the direct effect of the treatment is *negative*: $\hat{d} = -1$, even though the true direct effect is $d = +1$. The Baron-Kenny procedure is finding mediation where there is none and producing an estimate of the direct treatment effect that is wrongly signed.[3]

We might be sanguine about these results if they followed from a scenario that was unfavorable to nonexperimental mediation analysis, but they do not. On the contrary, the example provided here is more favorable to nonexperimental mediation analysis than any real-world application is likely to be. X, M, and Y have all been measured perfectly. There are no problems of reciprocal feedback from Y to X or M. The treatment effect is unambiguous: it is 1.0 not just on average but for every subject. It is completely unmediated by M—again, for every subject in the sample. And in spite of all of this, the Baron-Kenny estimates are incorrect.

To determine how frequently both treatments and mediators are manipulated in mediation analyses, we content-analyzed a random sample of 50 social psychology articles published in 2007 that cited Baron and Kenny (1986). Four of these articles did not use the Baron-Kenny method; Figure 1 shows how the method was applied in the other 46 articles. In a majority of these articles, mediation analysis involved manipulation of neither the independent variable nor the mediator. And in only one article—Bolger and Amarel (2007), to which we shall return—did the authors experimentally manipulate the mediators. (More information about the sample of articles appears in the online appendix.) To judge by this analysis, the Baron-Kenny procedure is almost always used in a way that is likely to exaggerate mediation effects.

The discussion in this section should make clear that the solution to bias in nonexperimental mediation analyses is not another nonexperimental mediation analysis. The problem is that the error terms in the models of the mediator and the dependent variable are likely to covary, and this is not a problem that we are likely to solve by controlling for more variables, measuring them more accurately, or using newer nonexperimental methods. To calculate unbiased estimates of mediation effects, we need to look to experimental methods.

Experimental Estimation of Mediation Effects

Mediation effects can be accurately analyzed with two-arm experiments in which each arm contains a sub-experiment. Subjects should be randomly assigned to one of the two arms. In the first arm, only the treatment is randomly assigned, and the mediator need not be measured. In the second arm, the treatment and all of the mediators are randomly assigned. This is not the only experimental design that can be used to accurately estimate mediation effects—indeed, a whole class of "two-study formats" (Taylor & Fiske, 1981) has been proposed—but it has much to recommend it.[4]

The second arm of the experiment has what Spencer et al. (2005, p. 846) term an "experimental-causal-chain" design.[5] To calculate the overall effect of the treatment on the dependent variable (c), we use the data from the first arm to estimate Equation 2. To calculate the direct effect of the treatment on the dependent variable (d), we use the data from the second arm to estimate a regression model in which the dependent variable is a function of the treatment and the potential mediators. (If there is only one potential mediator, this regression model is Equation 3.) The estimated mediation effect is $\hat{c} - \hat{d}$, where $\hat{c}$ and $\hat{d}$ are the estimates of c and d. This estimate is unbiased in expectation.

An example will clarify the approach. Suppose we know that hand-washing reduces the incidence of disease (e.g., Larson, 1988) and want to know whether the effect is mediated by the number of germs on subjects' hands. Following the form of Equation 3, our model is

$$\textit{diseases} = i_3 + d(\textit{hand-washing}) + b_1(\textit{germs}) + e_3, \qquad (6)$$

where *diseases* is a count of the number of diseases contracted by each subject in a period following the experiment, *hand-washing* is a binary variable indicating whether subjects washed their hands, *germs* is the number of germs on each subject's hands at the end of the experiment, i_3 is an intercept, d is the direct effect of hand-washing, and b_1 is the effect of germs. To estimate d and b_1, we conduct an experiment in which some subjects are assigned to wash their hands

$(X = 1)$ while others are not $(X = 0)$. Subjects assigned to wash their hands are further assigned to wash as they usually would $(M = 0)$ or to wash with a solution that adds germs to their hands $(M = 1)$. In all, there are three conditions: no hand-washing, hand-washing without added germs, and hand-washing with added germs. All hands are examined to determine germ counts, and afterward, all subjects are observed to determine the number of diseases that they contract.

This design is sufficient to permit accurate estimation of a mediation effect. By comparing subjects who do not wash their hands to those who wash their hands with ordinary water, we can recover an estimate of *c*, the total effect of hand-washing in ordinary circumstances. By using data from all subjects, we can use linear regression to estimate Equation 6, thereby recovering an estimate of *d*, the direct effect of hand-washing. In expectation, these estimates will be accurate, and if their difference $\left(\hat{c} - \hat{d}\right)$ is significantly different from zero, we will have strong evidence that germs mediate the effect of hand-washing.

The foregoing discussion says nothing about the more intractable problem of multiple mediators. Consider what happens if there is a second mediator of *hand-washing* in our model of *diseases*: hand-to-hand contact with others. In this case, we have

$$\begin{aligned}
germs &= i_{11} + a_1(hand\text{-}washing) + e_{11}, \\
contact &= i_{12} + a_2(hand\text{-}washing) + e_{12}, \\
diseases &= i_2 + c(hand\text{-}washing) + e_2, \text{ and} \\
diseases &= i_3 + d(hand\text{-}washing) + b_1(germs) + b_2(contact) + e_3,
\end{aligned}$$

where a_1 and a_2 are the effects of hand-washing on germs and hand-to-hand contact with others, and b_1 and b_2 are the effects of germs and hand-to-hand contact on the incidence of disease. If we manipulate hand-to-hand contact and include it in our model, the procedure is substantially unchanged. Our estimate of $c - d$ will still be accurate in expectation, and it will represent the cumulative effect of mediation by both germs and hand-to-hand contact.

If we neglect to include *contact* in our model of *diseases*, our ability to analyze mediation will be limited. Our estimator of b_1, the effect of germs on contraction of diseases, will still be unbiased, and we will be able to use it to estimate $a_1 b_1$, the extent to which the effect of hand-washing is mediated by the presence of germs on hands. But our estimator of d, the direct effect of hand-washing on contraction of diseases, will be biased: $E\left[\hat{d}\right] = d + a_2 b_2$. That is, our estimate of d will (in expectation) equal the true direct effect of hand-washing plus the part of its effect that is mediated by hand-to-hand contact with others. (Detailed calculations appear in the online appendix.) This may be a useful quantity to know, but because it is not d, our estimate of $c - d$ will not be accurate, and we will not be able to gauge the cumulative extent of mediation by germs and hand-to-hand contact. The lesson is general: failing to include one mediator in the analysis does not bar us from estimating the effects of other mediators, but it does bar us from estimating the cumulative extent of mediation. Uncertainty about multiple mediators is a fundamental limitation, and it makes the enterprise of learning about mediation a slow and difficult process.

To summarize, the virtue of this approach is its simplicity: it is a straightforward extension of the randomized experiments that social psychologists have long used to estimate treatment effects. So long as we can directly manipulate the hypothesized mediators, it will produce unbiased estimates of mediation effects. But note two difficulties with the procedure. One is that a *complete* mediation explanation—a complete accounting of the ways in which a treatment effect is mediated—requires that we know and simultaneously manipulate all of the mediators. Providing complete mediation explanations is a research agenda, not a task that can reasonably be assumed in a single article. And perhaps just as daunting is the requirement that all mediators be manipulated rather than merely observed. As a practical matter, many mediators are beyond direct experimental control. (See Spencer et al., 2005, for extensive discussion of this point.) But in these cases, too, we can improve on the Baron-Kenny approach and related approaches. The key in such cases is to manipulate variables that influence the mediators but do not directly affect the dependent variables. This is the method of instrumental variables, to which we turn now.

Instrumental-Variables Estimation of Mediation Effects

To produce unbiased estimates of mediation effects without directly manipulating mediators, we turn to the method of instrumental variables. It has been used to study mediation in epidemiology (Greenland, 2000), economics (Angrist & Krueger, 2001), and social policy analysis (Gennetian, Morris, Bos, & Bloom, 2005). But it is rare in psychology, where applications of the method to mediation analysis are almost nonexistent.

In an aside, Baron and Kenny (1986, p. 1177) mention that instrumental variables can be used to resolve issues raised by causal feedback in mediation chains. They can also be used to resolve issues raised by measurement error in the mediators. But even in the absence of reciprocal feedback and measurement error, the Baron-Kenny method and related methods are likely to produce biased estimates of mediation effects. Our contribution is to show explicitly how the instrumental-variables approach can be used to study mediation and why it solves the problem of bias to which nonexperimental methods are prone.[6]

Consider a scenario simpler than the ones presented in the previous sections. We have an outcome of interest, Y, and an explanatory variable, M. The regression equation is

$$Y = i + bM + e, \tag{7}$$

where b is the unknown true effect of M on Y, and e, the error term, represents all other variables that affect Y. If we could randomly assign different values of M to different subjects, we could ensure that M would be statistically independent of e, and the linear regression (i.e., ordinary least squares) estimator of b would therefore be unbiased. A researcher who does not manipulate M risks drawing biased inferences about M in the event that M is correlated with e.

A way to accurately estimate the effect of M without directly manipulating it is to find an instrument for it: a variable that is correlated with M but uncorrelated with e. Suppose that a third variable, Z, satisfies these criteria. To see how it helps us to estimate b, note that the covariance of Z and Y is in part a function of b:

$$\begin{aligned}\mathrm{cov}(Z, Y) &= \mathrm{cov}(Z, i + bM + e) \\ &= b \cdot \mathrm{cov}(Z, M) + \mathrm{cov}(Z, e).\end{aligned}$$

If $\mathrm{cov}(Z, e) = 0$, as would be expected if Z is a valid instrument,

$$\mathrm{cov}(Z, Y) = b \cdot \mathrm{cov}(Z, M),$$

which implies

$$b = \frac{\mathrm{cov}(Z, Y)}{\mathrm{cov}(Z, M)}. \tag{8}$$

We can use the data from our sample to estimate the covariances in the last equation: $\widehat{\mathrm{cov}}(Z, Y)$ is the estimate of $\mathrm{cov}(Z, Y)$, and $\widehat{\mathrm{cov}}(Z, M)$ is the estimate of $\mathrm{cov}(Z, M)$.[7] The instrumental-variables estimator of $\hat{b}$ is then $\frac{\widehat{\mathrm{cov}}(Z,Y)}{\widehat{\mathrm{cov}}(Z,M)}$. As sample size grows, this quantity converges on b: it solves the problem of endogeneity in Equation 7.[8]

This discussion presumes that Z is correlated with M but uncorrelated with e. Finding a variable that satisfies the first condition is often straightforward: whatever M is, it is usually not hard to think of a variable that is correlated with it, and once we have measured this new variable, testing the correlation is trivial. But satisfying the second condition is more difficult. Because e is unobservable, we can never directly test whether it is uncorrelated with Z. Worse, almost every variable that is correlated with M is likely to be correlated with other factors that affect Y, and thus likely to be correlated with e.

Fortunately, a familiar class of variables meets both conditions: randomly assigned treatments whose effect on Y is completely mediated by M. If Z is a randomly assigned treatment and we are confident that it only affects Y through M, we can use it as an instrument with which to estimate the effect of M. This is what we do when we use *encouragement designs*, in which we randomly encourage some subjects to expose themselves to the stimulus (Holland, 1988; see also Angrist et al., 1996; Jo, 2002; Little & Yau, 1998). In encouragement-design experiments, Z is the randomly assigned encouragement and M is actual exposure to the stimulus whose effect we want to know. For example, Z may be encouragement to study for a test and M may be the amount of time spent studying, as in Powers and Swinton (1984). So long as we can be confident that encouragement *per se* does not affect Y except through M, we can use Z as an instrument for M.

Of course, most of our interest in mediation lies not with situations represented by Equation 7 but with situations represented by Equation 3, in which we are unsure of the extent to which M mediates a randomly assigned treatment, X. Even if M is not randomly assigned, we can accurately estimate its effect if we have an instrument for it that does not appear in the model of Y (Hirano, Imbens, Rubin, & Zhou, 2000). For example, suppose that randomly assigned Z is an instrument for observed M in Equation 3. Our model of M includes both X and Z:

$$M = i_1 + a_1X + a_2Z + e_1. \tag{9}$$

(We include X in Equation 9 because it is, like Z, an exogenous variable: values of X are assigned independently of everything else that affects Y, and X is therefore uncorrelated with e_3 in Equation 3. All available exogenous variables should appear in the right-hand side of regression equations for M, whether or not we are using them as instruments for M.) Using X as an additional predictor slightly changes one of the requirements that must be met if Z is to be an instrument for M. It is no longer necessary or sufficient for Z and M to be correlated; instead, they must now be *partially* correlated: the requirement is $a_2 \neq 0$.

Although instrumental-variables estimation *per se* has not to our knowledge been used to estimate mediation effects in psychology (and a recent exegesis by MacKinnon, 2008 offers no examples), many social psychologists have used clever experiments to devise what are in effect instrumental variables for their mediators. For example, Smith (1982) hypothesizes that the effects of factual beliefs on attitudes are mediated by attributions. In his experiment, he relates a story in which a driver hits a child with his car. Factual beliefs are manipulated by randomly assigning subjects to hear that the driver did or did not keep his brakes in good condition. To test the mediating effect of attributions of responsibility on subjects' attitudes toward the driver, subjects are randomly assigned to hear that either the driver's or the child's carelessness was responsible for the accident. Zanna and Cooper (1974) hypothesize that attitude-behavior conflict produces an aversive state of arousal, which in turn produces attitude change. They cannot directly manipulate levels of aversive arousal, so they use an instrument to affect it indirectly: subjects are asked to swallow a pill and are randomly assigned to hear that it will have no effect, make them feel tense, or make them feel relaxed. Bolger and Amarel (2007) note from previous research (Bolger, Zuckerman, & Kessler, 2000) that social support is more likely to reduce recipients' stress when it is "invisible," i.e., not perceived as intended support. They hypothesize that the effect of support visibility on stress reduction is mediated by recipients' sense of efficacy: when people in stressful situations receive support that they perceive as an attempt to help, their sense of efficacy does not increase; but when they receive "invisible" support, it does. Bolger and Amarel cannot directly manipulate levels of efficacy among their subjects; instead, they use carefully scripted statements by confederates to indirectly manipulate subjects' sense of efficacy.

Each of these studies uses randomly generated interventions to create instrumental variables. But none of them use instrumental-variables estimation, and as a result, none of them properly estimate the effects of the mediators. For example, Zanna and Cooper (1974) measure arousal by asking subjects to report how they feel on a 31-point scale ranging from "calm" to "tense." Analysis of these self-reports shows that telling subjects how the pill will make them feel (the instrument) changes how they actually feel (the mediator). But instead of conducting an instrumental-variables analysis to estimate the effect of the mediator on attitudes (the dependent variable), Zanna and Cooper use ordinary regression to gauge the effect of the instrument on the dependent variable. This intent-to-treat effect may be informative, but it is not the same as the causal effect of the mediator, and this type of analysis can give a misleading impression of the extent of mediation (MacKinnon, 2008, p. 355). By contrast, instrumental-variables analyses do tell us about the effects of the mediator, and they thereby help us determine whether mediation is occurring.

For concreteness, reconsider the hand-washing example set forth in the previous section. This time, suppose that we can measure but cannot manipulate the numbers of germs on subjects' hands. We need an instrument for the germs-on-hands variable. A randomly introduced encouragement to use soap may suffice: for example, a poster by the sink that stresses the inadequacy of washing with water alone. If the presence of the poster is correlated with the number of germs on hands, and if it is statistically independent of unobserved factors that affect the incidence of disease, it is an instrument for the germs-on-hands variable.

Our model remains $diseases = i_3 + d(hand\text{-}washing) + b(germs) + e_3$. Because the factors other than hand-washing that affect *germs* are also likely to affect *diseases*, the Baron-Kenny procedure is likely to yield inaccurate estimates of *b* and *d*, and thus to mislead us about the extent to which germs-on-hands mediates the effect of hand-washing. But if we randomly assign the presence of the poster so that it is independent of hand-washing and all other variables that affect *diseases*, we create an instrument, *poster*, that we can use to accurately estimate the extent of mediation. Following the logic of Equation 8, we have

$$\begin{aligned} \operatorname{cov}(poster, diseases) &= \operatorname{cov}(poster, (i_3 + d(hand\text{-}washing) + b(germs) + e_3)) \\ &= b \cdot \operatorname{cov}(poster, germs) \\ \Rightarrow b &= \frac{\operatorname{cov}(poster, diseases)}{\operatorname{cov}(poster, germs)} \end{aligned}$$

and

$$\begin{aligned}\operatorname{cov}(\textit{hand-washing}, \textit{diseases}) &= \operatorname{cov}(\textit{hand-washing}, (i_3 + d(\textit{hand-washing}) + b(\textit{germs}) + e_3))\\ &= d + b \cdot \operatorname{cov}(\textit{hand-washing}, \textit{germs})\\ \Rightarrow d &= \operatorname{cov}(\textit{hand-washing}, \textit{diseases}) - b \cdot \operatorname{cov}(\textit{hand-washing}, \textit{germs}).\end{aligned}$$

The instrumental-variables estimator of b is $\hat{b} = \frac{\widehat{\operatorname{cov}}(\textit{poster},\textit{diseases})}{\widehat{\operatorname{cov}}(\textit{poster},\textit{germs})}$, and the instrumental-variables estimator of d is $\hat{d} = \widehat{\operatorname{cov}}(\textit{hand-washing}, \textit{diseases}) - \hat{b} \cdot \widehat{\operatorname{cov}}(\textit{hand-washing}, \textit{germs})$. Unlike the Baron-Kenny estimators of b and d, these estimators will converge to the correct values of b and d as sample size grows.

In this example, the instrumental-variables approach works only if *poster* is systematically related to *germs*. If the correlation between the variables is zero—as it will be in expectation if the random assignment of posters has no effect on the numbers of germs on subjects' hands—the encouragement has failed, and we do not really have an instrument for *germs*. Our model is *unidentified*: we cannot estimate its parameters with the data that we have and the assumptions that we are willing to make. We could solve the problem by making a new assumption; for example, assuming a particular value for b would let us compute a value of d. But there seems no way *a priori* to justify the assumption of any exact value for b. This is just one instance of the general identification problem: do our data and our assumptions permit us to learn what we want to know? It is less a statistical problem than a logical one (Manski, 1999)—a problem that social scientists have been grappling with since the 1940s, particularly in discussions of instrumental-variables estimation (e.g., Haavelmo, 1944; Koopmans, 1949).

We often hypothesize more than one mediator for any treatment effect that we want to explain. The instrumental-variables approach extends seamlessly to these situations, provided that we satisfy the *rank condition* (Koopmans, 1949). There are two components to the condition. First, we must have at least as many instruments as mediators. For example, if our model is

$$Y = i + dX + b_1 M_1 + b_2 M_2 + e \tag{10}$$

where M_1 and M_2 are mediators of Y and b_1 and b_2 are the effects of these mediators, the first component of the rank condition states that we must have at least two instruments to identify the effects of the mediators. Second, although each instrument may affect all of the mediators, each must affect the mediators in a different way. For example, if

$$M_1 = i_1 + a_{11}X + a_{12}Z_1 + a_{13}Z_2 + e_1, \text{ and} \tag{11}$$

$$M_2 = i_2 + a_{21}X + a_{22}Z_1 + a_{23}Z_2 + e_2, \tag{12}$$

where Z_1 and Z_2 are the instruments, the model of Equation 10 will be unidentified if $a_{12}/a_{13} = a_{22}/a_{23}$.[9]

The precision of instrumental-variables estimates is greatest when each instrument isolates a different mediator. For example, in Equations 11 and 12, the estimates will be most precise (that is, have the smallest standard errors) when Z_1 chiefly affects M_1 and Z_2 chiefly affects M_2. Experimental design can help to ensure that instruments isolate different mediators. This is exactly what Sheets and Braver (1999) do in their study of workplace harassment. Their hypothesis is that the effect of a harasser's status within an organization on perceptions of harassment by the "recipient" are mediated by both the harasser's social dominance and his power to influence the recipient's career. Social dominance and power to influence careers are abstractions that probably cannot be directly manipulated, but Sheets and Braver devise clever indirect manipulations. To manipulate social dominance, they randomly assign some subjects to read that the harasser is an attorney at a law firm while others read that he is an attorney at a law firm who twice failed to set up his own legal practice. To manipulate power to influence the recipient's career, subjects are randomly assigned to hear that the harasser is an attorney at either the recipient's firm or a different firm. Although manipulations like these are indirect, they are likely to affect the targeted mediators, to not affect other influences on the dependent variable, and thus to permit accurate estimation of mediation effects.

To make the multiple-instrument approach concrete, consider an example inspired by the classic postwar studies about the formation of friendships. These studies show that people who live closer to each other are more likely to become friends (e.g., Festinger, Schachter, & Back, 1950, ch. 4). To explain why residential proximity has this effect, two hypotheses are advanced. One is that familiarity mediates the effect of proximity: living close to others causes them to seem familiar to us, and familiarity in turn promotes friendship. A second, compatible hypothesis is that similarity mediates the effect of proximity: living close to others affords us the opportunity to discover what we have in common with them, and this discovered similarity promotes friendship (Newcomb, 1961).

Suppose that the structure of undergraduate residences affects the probability that undergraduates will befriend other randomly selected students from the same residences: the more rooms per dormitory floor (and thus, the more proximate one is to others), the more likely one is to make friends. Let *proximity* be an indicator of randomly-assigned residential proximity to others: $proximity = 1$ if a college freshman lives in a building containing 100 single rooms spread over two floors, and $proximity = 0$ if he lives in a building containing 100 single rooms spread over 10 floors. Let *friend*, the dependent variable, indicate the extent to which each subject in the study considers one other, randomly selected person from his dormitory a friend.[10] We further hypothesize that the effect of proximity on friendship is mediated by both familiarity and similarity. Let *familiarity* indicate the extent to which this person seems familiar to the subject, and let *similarity* be the extent to which this person seems similar to the subject. The relevant equations have the forms of Equations 2 and 3:

$$friend = i_2 + c(proximity) + e_2 \tag{13}$$

$$friend = i_3 + d(proximity) + b_1(familiarity) + b_2(similarity) + e_3 \tag{14}$$

If we use the Baron-Kenny method to estimate Equation 14, our estimates of d, b_1, and b_2 are likely to be biased in favor of finding mediation. If we could randomly assign *familiarity* and *similarity*, we could use the procedure described in the previous section to accurately estimate those coefficients. But there is no obvious way to randomize *familiarity* and *similarity*. We need instrumental variables.

Let Z_{fam} indicate random assignment of some subjects to dormitories that either encourage or discourage familiarity among residents. For example, $Z_{fam} = 1$ might indicate assignment to a dormitory with a common lounge and laundry room; $Z_{fam} = 0$ might indicate assignment to a dormitory without any common spaces. Z_{fam} is a plausible instrument for *familiarity*: it is likely to affect the extent to which subjects find other people familiar, but because it is randomly assigned, it will be uncorrelated with the other factors that affect the development of friendship, and thus with e_3. Similarly, let Z_{sim} indicate random assignment of subjects to live near people who have similar background attributes. Z_{sim} is a plausible instrument for *similarity*: it is likely to affect the extent to which subjects discover similarity in others, but because it is randomly assigned, it will be uncorrelated in expectation with the other factors that affect the development of friendship, and thus with e_3. Figure 2 indicates the relations between the variables.

Randomly assign some subjects to an arm of the experiment in which only *proximity* is randomized. Assign other subjects to an arm in which *proximity*, Z_{fam}, and Z_{sim} are all randomized. This second arm is a sub-experiment with a $2 \times 2 \times 2$ design: each subject in this arm is assigned to one of two dormitories, to high- or low-familiarity encouragement, and to high- or low-similarity neighbors.[11] With the data from the first arm, in which only *proximity* is randomized, we can use linear regression to estimate Equation 13, thereby gaining an estimate of c, the total effect of proximity, that is accurate in expectation. With the data from the second arm, we can use our instrumental variables to gain estimates of d, the direct effect of proximity, and b_1 and b_2, the effects of familiarity and similarity, that are also accurate in expectation.

To see the difference between Baron-Kenny estimates and instrumental-variables estimates, consider some simulated data. We assume that 200 subjects participate in the experiment just described. We also assume

$$familiarity = 1.0(proximity) + 1.0(Z_{fam}) + e_{fam} \text{ and}$$

$$similarity = 1.0(proximity) + 1.0(Z_{sim}) + e_{sim},$$

where e_{fam} and e_{sim} are normally distributed errors with mean 0 and variance 1. Note that each instrument isolates its mediator: Z_{fam} only affects *familiarity*, and Z_{sim} only affects *similarity*. This ensures that the rank condition will be satisfied, and it increases the precision of the instrumental-variables estimates.

Next, assume that the outcome variable, *friend*, is a continuous variable for which higher values indicate that the subject feels more friendship toward the person about whom experimenters are asking. The direct effect of *proximity* and the mediating effects of *familiarity* and *similarity* are given by

$$friend = 10.0(proximity) + 10.0(familiarity) + 5.0(similarity) + e_3. \tag{15}$$

Note the correspondence between Equation 14 and Equation 15. The direct effect of *proximity*, 10 points, corresponds to d in Equation 14. And the coefficients on *familiarity* and *similarity*, 10 and 5, correspond to b_1 and b_2 in Equation 14. The total effect of *proximity* is $10 + (1 \cdot 10) + (1 \cdot 5) = 25$ points on the continuous *friend* scale; this corresponds to c in Equation 13.

Other details are provided in the online appendix, but this is all that is required to simulate an experiment. The estimates yielded by the Baron-Kenny procedure are $\hat{c} = 25.64$ (s.e. = 2.44), $\hat{d} = 4.89$ (s.e. = .45), $\widehat{b_1} = 12.41$ (s.e. = .17), and $\widehat{b_2} = 7.48$ (s.e. = .17). Thanks to the random assignment of X, our estimate of c, the total effect of *proximity*, is very close to the true total effect, and the difference between the estimate and the true parameter value is statistically insignificant ($p = .79$). But the Baron-Kenny estimate of d, the direct effect of *proximity*, is less than half of what it should be. And the estimates of b_1 and b_2, which indicate the extent to which the effect of *proximity* is mediated by *familiarity* and *similarity*, are inflated. (All three of these estimates differ from their true parameter values at $p < .001$. Note that the estimates are far from the true values even though $R^2 = .98$ for the fitted model.) There are no such problems with the instrumental-variables estimates: $\hat{d} = 9.41$ (s.e. = 1.11, $p = .60$), $\widehat{b_1} = 10.48$ (s.e. = .55, $p = .38$), and $\widehat{b_2} = 5.11$ (s.e. = .64, $p = .86$). While the standard errors are larger for these estimates—a necessary consequence of manipulating the mediators indirectly—the estimates themselves are accurate.

Figure 3 reveals the same pattern when the experiment is simulated 1,000 times. The Baron-Kenny estimates, reported in the top row, substantially overstate the extent of mediation. The problem is not just that these estimates are too big on average; rather, *every* estimate is far from the true value of its parameter. Although the true value of d is 10, the largest Baron-Kenny estimate is only 6.57; although the true value of b_1 is 10, the smallest Baron-Kenny estimate is 11.85; and although the true value of b_2 is 5, the smallest Baron-Kenny estimate is 6.78. By contrast, the instrumental-variables estimates, reported in the bottom row, are evenly distributed around the true parameter values.

Mediation, Experiments, and Causal Inference in Psychology

We are not the first to argue that nonexperimental methods of mediation analysis are deeply problematic (e.g., Stone-Romero & Rosopa, 2008; Robins & Greenland, 1992; see also Aronson, Wilson, & Brewer, 1998, p. 105). The argument has been met in the past by several responses, one of which deserves special attention: requiring manipulation “precludes the study of many variables where manipulation is not ethically or practically possible" (Kenny, 2008, p. 356; see also James, 2008). This objection fails on three counts.

First, the objection follows from a misunderstanding of the argument. No one, so far as we know, maintains that unmanipulable variables should not be studied or that causal inferences should be drawn only from experiments; certainly, these are not our positions. The issue lies instead with the accuracy of nonexperimental inferences and the degree of confidence that we should place in them. So-called "natural experiments" occasionally produce instrumental variables, and analyses of these events may justify moderately strong inferences (Angrist & Krueger, 2001; Rosenzweig & Wolpin, 2000). And when effects are dramatic, we can reasonably infer their existence without either real or natural experiments (Freedman, 2005, p. 17; Smith & Pell, 2003). But in the absence of natural experiments, dramatic effects, or detailed theory about data-generating processes—that is, in almost all situations that social scientists examine—nonexperimental studies are likely to produce biased estimates of causal effects and to justify only very weak causal inferences.

Second, those who raise the objection treat it as a point in favor of nonexperimental methods of mediation analysis, which it is not. For example, James (2008, p. 361) writes "Well, then let's not engage in causal inferences about the effects of religion on attitudes, the ethics of the Enron executives. . . [or] what makes terrorists engage in suicide attacks." James is sarcastic, but he has stumbled upon a serious point. Religion, ethics, and the causes of terrorism are worth studying, but causal inferences about them are fragile and likely to remain so. All the nonexperimental studies of these topics that have ever been conducted do not mean as much as a small set of well-designed experiments would. Kenny (2008, p. 356) objects that this line of reasoning "artificially limits" the production of knowledge. But as we have argued, the real artifice lies in tacitly assuming, as users of nonexperimental methods of mediation analysis do, that Nature runs readily interpretable experiments for us. Some causal effects are inherently more difficult to learn about than others, and "some variables cannot be manipulated" is no defense of nonexperimental mediation analyses.

Third, the objection is unduly pessimistic. In many cases, we can learn about relevant effects by redefining the variable of interest so that it is manipulable: not race or gender, but the information about race or gender provided on a job application; not "ethics," but the circumstances under which different ethical intuitions are primed. To be sure, there are cases in which this approach will not suffice and in which variables of interest cannot be manipulated. That said, the history of social psychology is in large part a history of ingenious manipulations (Aronson et al., 1998; Jones, 1998; Rodrigues & Levine, 1999). And mediators need not be directly manipulated; as we have shown, indirect manipulation is sufficient to produce unbiased estimates of mediation effects. Many variables that cannot be directly manipulated can be indirectly manipulated, and the examples in the previous section show that psychologists have been indirectly manipulating mediators for decades. Experiments are difficult, and so is causal inference—but in light of the discipline's history of creative manipulation, we see more cause for optimism than for despair.

Conclusion

Few will dispute that psychological research on mediation has brought a new level of sophistication to the way in which social scientists think about causal pathways. Criticisms of mediation analysis attest to its influence even as they call attention to its limitations. But most previous criticisms have neglected the fundamental problem of producing accurate estimates of mediation effects when mediators are not manipulated. Unless mediators are manipulated, estimates of mediation effects are likely to be biased even in large-sample studies that are unaffected by measurement error, uncertainty about the direction of causal effects, or other problems that have been the focus of methodological mediation research for the past two decades.

Bias is a grave problem, but graver still is the erroneous impression that previous articles on mediation analysis have instilled in many readers: the impression that trustworthy mediation analysis is easy, or at least no more difficult than running an additional regression. Conventional methods of mediation analysis do not require the researcher to undertake additional experiments. In reality, trustworthy mediation analysis almost always requires experimentation. Without experimentation, there is generally no way to ensure that estimates of mediation effects are unbiased.

Moreover, the kind of experimentation required for the study of mediation is more difficult than the kind that is required to demonstrate a causal effect. It requires not just manipulation (direct or indirect) of mediators, but simultaneous manipulation of both treatments and mediators. And complete explanation of the ways in which effects are transmitted from independent to dependent variables typically cannot be achieved by a single study, no matter how careful. As Kazdin (2007, p. 11) has it, understanding mediators "is not a matter of one study but is a matter of creeping up on the process that draws on a series of projects."

Because experimental approaches produce unbiased estimates of mediation effects, they stand to produce firmer knowledge of mediation than we have yet acquired. Less obviously, they also stand to teach us more about nonexperimental approaches to mediation analysis. Precisely because experimental estimates of mediation effects are unbiased, they are benchmarks against which we can gauge the accuracy of nonexperimental estimates, including those produced by the Baron-Kenny approach. The use of experiments as benchmarks against which to test nonexperimental methods of data analysis was pioneered by LaLonde (1986), but it has yet to be applied to mediation analysis, because hypotheses about mediators have yet to be tested experimentally with enough precision to provide a reliable benchmark. We hope that this will soon change.

We have described two experimental designs that suffice to produce accurate estimates of causal effects. One requires direct manipulation of mediators. Another permits indirect manipulation via instrumental variables, a technique that is increasingly popular in other social sciences and that we hope will become common in social psychology. Whether one uses the direct or the indirect approach, the main challenge is the same: it is to devise experimental manipulations that principally affect one mediator without affecting others. This is what some social psychologists have been doing for decades, and a practice to which the field should return.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444-55.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives, 15,* 69-86.

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology: Vol. 1* (4th ed., pp. 99-142). New York: McGraw-Hill.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173-82.

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage Publications.

Blau, P. M., & Duncan, O. D. (1967). *The American occupational structure*. New York: John Wiley & Sons.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8,* 225-46.

Bolger, N., & Amarel, D. (2007). Effects of social support visibility on adjustment to stress: Experimental evidence. *Journal of Personality and Social Psychology, 92,* 458-75.

Bolger, N., Zuckerman, A., & Kessler, R. C. (2000). Invisible support and adjustment to stress. *Journal of Personality and Social Psychology, 79,* 953-61.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association, 90,* 443-50.

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

Festinger, L., Schachter, S., & Back, K. (1950). *Social pressures in informal groups*. Stanford, CA: Stanford University Press.

Fisher, R. A. (1935). *The design of experiments*. 1st ed. Edinburgh: Oliver and Boyd.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58,* 21-29.

Freedman, D. A. (2005). *Statistical models: Theory and practice*. New York: Cambridge University Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gennetian, L. A., Morris, P. A., Bos, J. M., & Bloom, H. S. (2005). Constructing instrumental variables from experimental data to explore how treatments produce effects. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 75-114). New York: Russell Sage.

Greenland, S. (2000). Causal analysis in the health sciences. *Journal of the American Statistical Association, 95,* 286-89.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica, 12,* 1-115.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics, 1,* 69-88.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. *Sociological Methodology, 18,* 449-84.

Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* . Newbury Park, CA: Sage.

James, L. R. (1980). The unmeasured variables problem in path analysis. *Journal of Applied Psychology, 65,* 415-21.

James, L. R. (2008). On the path to mediation. *Organizational Research Methods, 11,* 359-63.

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

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics, 27,* 385-409.

Jones, E. E. (1998). Major developments in five decades of social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 3-57). New York: McGraw-Hill.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602-19.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3,* 1-27.

Kenny, D. A. (2008). Reflections on mediation. *Organizational Research Methods, 11,* 353-58.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 233-68). New York: McGraw-Hill.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica, 17,* 125-44.

Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *America Journal of Psychiatry, 158,* 848-56.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76,* 604-20.

Larson, E. (1988). A causal link between handwashing and risk of infection? Examination of the evidence. *Infection Control, 9,* 28-36.

Lazarsfeld, P. F. (1955). Interpretation of statistical relations as a research operation. In P. F. Lazarsfeld & M. Rosenberg (Eds.), *The language of social research: A reader in the methodology of social research* (pp. 115-25). Glencoe, IL: Free Press.

Little, R. J., & Yau, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3,* 147-59.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Lawrence Erlbaum Associates.

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

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

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

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research, 30,* 41-62.

Manski, C. F. (1999). *Identification problems in the social sciences*. 2nd ed. Cambridge, MA: Harvard University Press.

Mathieu, J. E., & Taylor, S. R. (2006). Clarifying conditions and decision points for mediational type inferences in organizational behavior. *Journal of Organizational Behavior, 27,* 1031-56.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12,* 23-44.

McDonald, R. P. (1997). Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research, 32,* 1-38.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89,* 852-863.

Newcomb, T. M. (1961). *The acquaintance process*. Austin, TX: Holt, Rinehart and Winston.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology, 76,* 266-78.

Quiñones-Vidal, E., López-Garcia, J. J., Peñaranda-Ortega, M., & Tortosa-Gil, F. (2004). The nature of social and personality psychology as reflected in *JPSP*, 1965-2000. *Journal of Personality and Social Psychology, 86,* 435-52.

R Development Core Team. (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3,* 143-55.

Rodrigues, A. & Levine, R. V. (Eds.). (1999). *Reflections on 100 years of experimental social psychology*. New York: Basic Books.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A, 147,* 656-66.

Rosenzweig, M. R., & Wolpin, K. I. (2000). Natural “natural experiments” in economics. *Journal of Economic Literature, 38,* 827-74.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31,* 161-70.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sheets, V. L., & Braver, S. L. (1999). Organizational status and perceived sexual harassment: Detecting the mediators of a null effect. *Personality and Social Psychology Bulletin, 25,* 1159-71.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7,* 422-45.

Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review, 2,* 218-26.

Smith, E. R. (1982). Beliefs, attributions, and evaluations: Nonhierarchical models of mediation in social cognition. *Journal of Personality and Social Psychology, 43,* 248-59.

Smith, G. C. S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *British Medical Journal, 327,* 1459-61.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89,* 845-851.

Stone-Romero, E. F., & Rosopa, P. J. (2008). The relative validity of inferences about mediation as a function of research design characteristics. *Organizational Research Methods, 11,* 326-52.

Sykes, R. E. (1983). Initial interaction between strangers and acquaintances: A multivariate analysis of factors affecting choice of communication partners. *Human Communication Research, 10,* 27-53.

Taylor, S. E., & Fiske, S. T. (1981). Getting inside the head: Methodologies for process analysis in attribution and social cognition. In J. Harvey, W. Ickes, & R. Kidd (Eds.), *New directions in attribution research: Vol. 3* . Hillsdale, NJ: Lawrence Erlbaum.

Wood, R. E., Goodman, J. S., Beckmann, N., & Cook, A. (2008). Mediation testing in management research: A review and proposals. *Organizational Research Methods, 11,* 270-95.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Zanna, M. P., & Cooper, J. (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology, 29,* 703-09.

Footnotes

1. To see why the mediated effect is ab, note that substituting Equation 1 into Equation 3 yields $Y = (bi_1 + i_3) + (ab + d)\,X + (be_1 + e_3)$. The total effect of X in this equation is given by $(ab + d)$. Since d is the coefficient on X in Equation 3, it is called the "direct" effect; the effect that is not "direct" is "mediated." For proof that $ab = c - d$ when there is only one mediator, see MacKinnon, Warsi, and Dwyer (1995).

2. Throughout this article, we assume that the Baron-Kenny procedure uses ordinary least squares (i.e., linear regression) estimators of the parameters in Equations 1 through 3.

3. We do not focus on statistical significance in this example because the number of cases represented by Table 2 is arbitrary. Without loss of generality, each row of the table can be assumed to represent not one case but n cases, where n is an arbitrarily large integer. For the purpose of the example, the important assumption is that each row of the table represents an equal number of cases.

4. Thorough discussion of the advantages of this design is beyond the scope of this paper, but two advantages deserve mention. Most two-study formats call for manipulation of one variable at a time: X is manipulated in one experiment and M is manipulated in another experiment. This design permits the estimation of mediation effects, but it does not permit analysis of interactions between X and M, as our design does. Moreover, our design eliminates the requirement of taking multiple samples from the same population, which must be faced by those who follow traditional two-study formats.

5. Spencer et al. state that such designs permit discovery of mediation but make it difficult to estimate the extent of mediation. This is true only in a limited sense. When the experiment in the second arm is paired with the more conventional experiment in the first arm, it is possible to accurately estimate the extent of mediation.

6. The method of instrumental variables receives expansive treatment in many textbooks, e.g., Freedman (2005, ch. 8). Our discussion here is narrower and devoted to making the connection to mediation.

7. $\widehat{\text{cov}}(Z, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(Y_i - \overline{Y}\right)$, where n is the number of subjects in the sample, Z_i and Y_i are the observations of Z and Y for subject i, and $\overline{Z}$ and $\overline{Y}$ are the sample means of Z and Y. Similarly, $\widehat{\text{cov}}(Z, M) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)\left(M_i - \overline{M}\right)$.

8. The instrumental-variables estimator is *consistent*: as sample size tends to infinity, instrumental-variables estimates of b converge to b. The estimator is biased in small samples, but the bias is minuscule when the instrument and the endogenous variable are highly correlated, as they are almost certain to be when the endogenous variable is an experimental treatment and the instrument is random assignment to the treatment or to encouragement to take the treatment. Even if the instrument and the endogenous variable are weakly correlated, the bias in the instrumental-variables estimator is necessarily smaller than the bias in the linear regression (i.e., ordinary least squares) estimator if the instrument and the error term in the model of the dependent variable do not covary. For more information about bias and consistency of instrumental-variables estimators, see Bound, Jaeger, and Baker (1995).

Some subjects will never take a treatment even if they are assigned to take it (Bloom, 1984). A treatment effect cannot be meaningfully estimated for such people. Consequently, the instrumental-variables estimator $\hat{b}$ estimates the "local average treatment effect" (LATE): the average effect of the treatment on subjects who can be induced by random assignment to take it. For more on the LATE and related estimands, see Gennetian et al. (2005) and Angrist et al. (1996).

9. Formally, the rank condition states that if $\mathbf{M}$ is the $n \times p$ matrix of mediators and $\mathbf{Z}$ is the $n \times q$ matrix of instruments, then d and $b_1, \ldots, b_n$ are identified if and only if the matrix $\mathbf{Z}'\mathbf{M}$ has full rank (Wooldridge, 2002, pp. 85-86).

10. The study can be extended to allow each subject to report the extent of his friendship with every other person in the study, but only at the expense of considerable notational complexity.

11. Sykes (1983) uses a similar design to study how people choose "communication partners."

Table 1

The Baron-Kenny Procedure May Lead to Biased Inferences

Fraction	Potential outcomes				Observed data		
of population	*M(1)*	*M(0)*	*Y(1)*	*Y(0)*	*X*	M_{obs}	Y_{obs}
1/6	3	2	11	10	0	2	10
1/6	3	2	11	10	1	3	11
1/6	3	3	13	12	0	3	12
1/6	3	3	13	12	1	3	13
1/6	4	3	15	14	0	3	14
1/6	4	3	15	14	1	4	15

Note. *Y(1)* is the set of values that the dependent variable would assume if all subjects were treated. *Y(0)* is the set of values that it would assume if no subjects were treated. *M(1)* and *M(0)* are the analogous values for the mediator. Inspection of these "potential outcomes" shows that the treatment effect is +1 for each subject in the sample and that this effect is entirely unmediated: $c = d = +1$. But the Baron-Kenny procedure indicates that *M* heavily mediates *X* and that the direct effect of *X* is *negative*. This table uses the example provided by Rubin (2005, p. 328).

Figure Captions

Figure 1. Content Analysis of 50 Articles Published in 2007 that Use the Baron-Kenny Method of Mediation Analysis. Fifty articles were sampled from the population of social psychology articles that cited Baron and Kenny (1986) and were published in 2007. Four of the articles did not use the Baron-Kenny method of mediation analysis. Of the other 46, 45 applied the method in a way that is likely to exaggerate mediation effects. (Further details about the content analysis appear in the online appendix.)

Figure 2. Causal Relations in the Residential Proximity Experiment. The experiment is designed to test whether the effect of residential proximity on the formation of friendships is mediated by perceptions of familiarity and similarity. Arrows indicate causal relationships between variables. Z_{fam} is an instrument that only affects *familiarity*, and Z_{sim} is an instrument that only affects *similarity*. (For simplicity, error terms are not displayed.)

Figure 3. Simulated Baron-Kenny Estimates versus Simulated Instrumental-Variables Estimates. The residential proximity experiment was simulated 1,000 times. Vertical lines in each panel indicate the true parameter values: $d = 10$, $b_1 = 10$, and $b_2 = 5$. For each simulation, d, b_1, and b_2 were estimated by the Baron-Kenny procedure and by instrumental variables. The Baron-Kenny estimates appear in the top row; the instrumental-variables estimates appear in the bottom row. The variation in the Baron-Kenny estimates is smaller, but on average, they are far from the true values of the parameters. By contrast, the histograms depicting the instrumental-variables estimates are centered on the true values of the parameters.

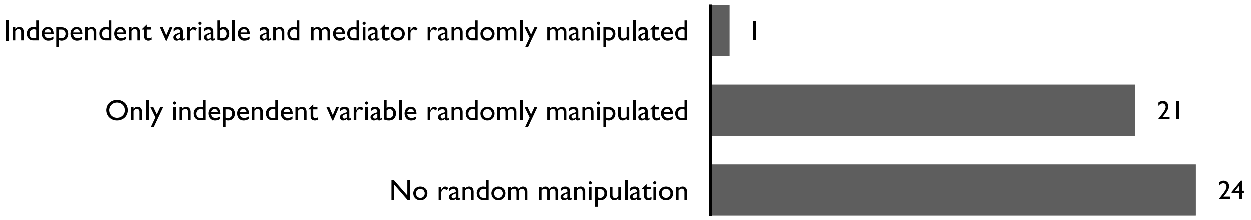
Independent variable and mediator randomly manipulated
1
Only independent variable randomly manipulated
21
No random manipulation
24

Figure 1.

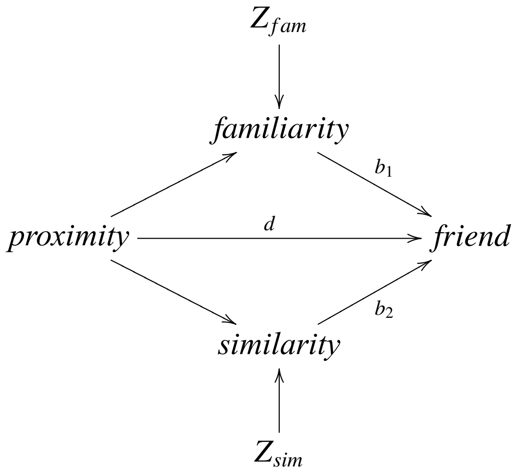

Figure 2.

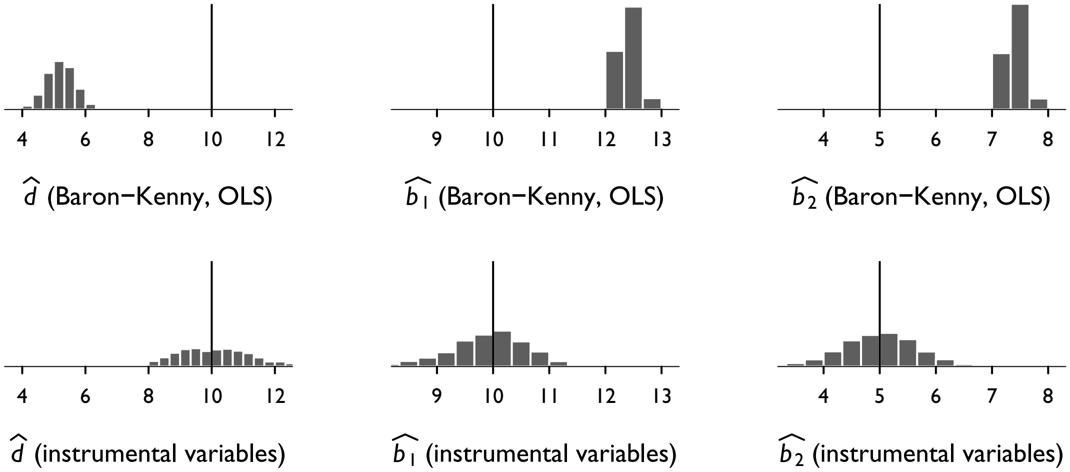
4
6
8
10
12
$\widehat{d}$ (Baron−Kenny, OLS)
9
10
11
12
13
$\widehat{b_1}$ (Baron−Kenny, OLS)
4
5
6
7
8
$\widehat{b_2}$ (Baron−Kenny, OLS)
4
6
8
10
12
$\widehat{d}$ (instrumental variables)
9
10
11
12
13
$\widehat{b_1}$ (instrumental variables)
4
5
6
7
8
$\widehat{b_2}$ (instrumental variables)

Figure 3.

Online Supplement: Appendix

Bias in Estimates of b and d

Consider the equation

$$Y = d\tilde{X} + b\tilde{M} + e_3, \tag{A1}$$

where $\tilde{X} = X - \overline{X}$, $\tilde{M} = M - \overline{M}$, and X, M, and e_3 are as defined on page 5. Biases in the Baron-Kenny (i.e., ordinary least squares, linear regression) estimators of d and b in Equation A1 are the same as the biases for the Baron-Kenny estimators of d and b in Equation 3. Following Muller, Judd, and Yzerbyt (2005), we use the mean-centered predictors of Equation A1, which make for easier interpretation and greatly simplify the calculations.

Let $\tilde{\mathbf{X}}$ be the design matrix $[\tilde{X}, \tilde{M}]$. The linear regression (ordinary least squares) estimators are

$$\hat{\beta} = \begin{bmatrix} \hat{d} \\ \hat{b} \end{bmatrix} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(d\tilde{X} + b\tilde{M} + e_3)$$

$$= \begin{bmatrix} d \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ b \end{bmatrix} + (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'e_3, \text{ where}$$

$$\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'e_3 = \begin{bmatrix} \frac{\sum {e_1}_i^2 \sum \tilde{X}_i {e_3}_i + a \sum \tilde{X}_i {e_1}_i \sum \tilde{X}_i {e_3}_i - \sum {e_1}_i \tilde{X}_i \sum {e_1}_i {e_3}_i - a \sum \tilde{X}_i^2 \sum {e_1}_i {e_3}_i}{\sum {e_1}_i^2 \sum \tilde{X}_i^2 - \sum^2 {e_1}_i \tilde{X}_i} \\ \\ \frac{\sum {e_1}_i {e_3}_i \sum \tilde{X}_i^2 - \sum {e_1}_i \tilde{X}_i \sum \tilde{X}_i {e_3}_i}{\sum {e_1}_i^2 \sum \tilde{X}_i^2 - \sum^2 {e_1}_i \tilde{X}_i} \end{bmatrix}.$$

Now,

$$\begin{aligned} E[\hat{b}] &= b + E\left[\frac{\sum e_{1i}e_{3i}\sum \tilde{X}_i^2 - \sum e_{1i}\tilde{X}_i \sum \tilde{X}_i e_{3i}}{\sum {e_1}_i^2 \sum \tilde{X}_i^2 - \sum^2 e_{1i}\tilde{X}_i}\right] \\ &= b + \frac{\operatorname{cov}(e_1, e_3)}{\operatorname{var}(e_1)}. \end{aligned}$$

And by the same logic,

$$E[\hat{d}] = d - E\left[\frac{a\sum e_{1i}e_{3i}}{\sum {e_1}_i^2}\right] = d - a \cdot \frac{\operatorname{cov}(e_1, e_3)}{\operatorname{var}(e_1)}.$$

Thus, estimating a Baron-and-Kenny type regression leads to bias in the estimates of both d and b whenever $\operatorname{cov}(e_1, e_3) \neq 0$.

Matrix Notation for Application of the Baron-Kenny Method to Table 2

The first equation is

$$\begin{array}{ccccccccc} M_{obs} & & i_1 & & a & X & & e_1 & \\ \begin{bmatrix} 2 \\ 3 \\ 3 \\ 3 \\ 3 \\ 4 \end{bmatrix} & = & \begin{bmatrix} 2.67 \\ 2.67 \\ 2.67 \\ 2.67 \\ 2.67 \\ 2.67 \end{bmatrix} & + & .67 & \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} & + & \begin{bmatrix} -.67 \\ -.33 \\ .33 \\ -.33 \\ .33 \\ .67 \end{bmatrix} & , \end{array}$$

the second equation is

$$\begin{array}{ccccccccc} Y_{obs} & & i_2 & & c & X & & e_1 & \\ \begin{bmatrix}10\\11\\12\\13\\14\\15\end{bmatrix} & = & \begin{bmatrix}12\\12\\12\\12\\12\\12\end{bmatrix} & + & 1 & \begin{bmatrix}0\\1\\0\\1\\0\\1\end{bmatrix} & + & \begin{bmatrix}-2\\-2\\0\\0\\2\\2\end{bmatrix} & , \end{array}$$

and the third equation is

$$\begin{array}{cccccccccccc} Y_{obs} & & i_3 & & b & M_{obs} & & d & X & & e_3 & \\ \begin{bmatrix}10\\11\\12\\13\\14\\15\end{bmatrix} & = & \begin{bmatrix}12\\12\\12\\12\\12\\12\end{bmatrix} & + & 0 & \begin{bmatrix}2\\3\\3\\3\\3\\4\end{bmatrix} & + & 1 & \begin{bmatrix}0\\1\\0\\1\\0\\1\end{bmatrix} & + & \begin{bmatrix}-2\\-2\\0\\0\\2\\2\end{bmatrix} & . \end{array}$$

Random Sample of Psychology Articles that Cite Baron and Kenny (1986)

A PsycINFO search produced a list of 168 psychology articles published in 2007 that cited Baron and Kenny (1986). From this population, we drew a random sample of 50. Four did not use the Baron and Kenny method of mediation analysis. The other 46 are listed in Table A1.

Estimation of Mediation Effects When Mediators Are Omitted from the Analysis

Page 14 considers a mediation analysis from which a mediator is omitted. In such cases, the estimate of *d*, the direct effect of the treatment, equals (in expectation) the true direct effect plus the part of the overall effect that is mediated by the omitted mediator. Formally, consider a treatment effect that is mediated by two variables:

$$\begin{aligned} M_1 &= i_1 + a_1 X + \epsilon_1 \\ M_2 &= i_2 + a_2 X + \epsilon_2 \\ Y &= i_3 + b_1 M_1 + b_2 M_2 + dX + \epsilon_3. \end{aligned}$$

where X is a treatment and M_1 and M_2 are mediators. Assume that X, M_1, and M_2 are randomly assigned such that $\text{cor}(X, \epsilon_3) = \text{cor}(M_1, \epsilon_3) = \text{cor}(M_2, \epsilon_3) = 0$ in expectation. If we omit M_2 from our analysis, we estimate

$$Y = i_3 + b_1 M_1 + dX + v_1$$

where $v_1 = b_2 M_2 + \epsilon_3 = b_2(i_2 + a_2 X + \epsilon_2) + \epsilon_3$. If we estimate this model by the Baron-Kenny method, we implicitly presume $X \perp\!\!\!\perp v_1$ and thus $\text{cor}(X, v_1) = 0$ in expectation. In expectation, the Baron-Kenny estimate of d will be $\hat{d} = \text{cov}(X, Y) - a_1 b_1$:

$$\begin{aligned} \text{cov}(X, Y) &= \text{cov}(X, i_3 + b_1 M_1 + dX + v_1) \\ &= a_1 b_1 + d \\ \Rightarrow d &= \text{cov}(X, Y) - a_1 b_1. \end{aligned}$$

In reality, v_1 is a function of X and is therefore not independent of it. The true value of d is $\text{cov}(X, Y) - a_1 b_1 - a_2 b_2$:

$$\begin{aligned} \text{cov}(X, Y) &= \text{cov}(X, i_3 + b_1 M_1 + dX + v_1) \\ &= a_1 b_1 + a_2 b_2 + d \\ \Rightarrow d &= \text{cov}(X, Y) - a_1 b_1 - a_2 b_2. \end{aligned}$$

We have $E\left[\hat{d}\right] - d = a_2 b_2$: the estimator of d is biased. This is not entirely a negative result—see page 14 for details—but it means that we cannot rely on the estimate of d to indicate the true direct effect of X on Y.

Propinquity Simulations

Assume that *proximity*, Z_{fam}, and Z_{sim} are all distributed binomially with $N = 200$ and $p = .5$. We have

$$familiarity = proximity + Z_{fam} + e_{fam} \text{ and}$$

$$similarity = proximity + Z_{sim} + e_{sim},$$

where $e_{fam} \sim N(0, 1)$ and $e_{sim} \sim N(0, 1)$. Finally,

$$friend = 10 \cdot proximity + 10 \cdot familiarity + 5 \cdot similarity + e_3$$

where $e_3 = \epsilon + 3e_{fam} + 3e_{sim}$ and $\epsilon \sim N(0, 2)$. We use the simulated values of *proximity*, *familiarity*, and *similarity* to compute the Baron-Kenny estimates. To compute the instrumental-variables estimates, we also use the simulated values of Z_{fam} and Z_{sim}. The simulations were conducted in `R` 2.6.1 (R Development Core Team, 2007) with these commands:

```
set.seed(2008)                     # seed the random-number generator
N  <- 500                          # number of subjects
d  <-  10                          # direct effect of proximity
b1 <-  10                          # effect of familiarity
b2 <-   5                          # effect of similarity
proximity   <- rbinom(N, 1, .5)    # 50% assigned to high-proximity residence
Z.fam       <- rbinom(N, 1, .5)    # 50% assigned to familiarity encouragement
Z.sim       <- rbinom(N, 1, .5)    # 50% assigned to similarity encouragement

e.fam       <- rnorm(N, mean=0, sd=1)
```

```
e.sim        <- rnorm(N, mean=0, sd=1)
familiarity <- Z.fam + proximity + e.fam
similarity  <- Z.sim + proximity + e.sim

e.3          <- rnorm(N, mean=0, sd=2) + 3*e.fam + 3*e.sim
friend       <- d*proximity + b1*familiarity + b2*similarity + e.3

summary(lm(friend ~ proximity))
summary(lm(friend ~ proximity + familiarity + similarity))
library(sem)
summary(tsls(friend ~     proximity + familiarity + similarity
                      , ~ proximity + Z.fam + Z.sim))
```

`lm(friend ~ proximity)` uses ordinary least squares to fit the regression of *friend* on *proximity*. As expected, the OLS estimate of the overall effect of *proximity* is very close to the true effect: $\hat{c}$ = 25.64 (s.e. = 2.44).

`lm(friend ~ proximity + familiarity + similarity)` uses OLS to fit the regression of *friend* on *proximity*, *familiarity*, and *similarity*. This is the final step of the Baron-Kenny procedure. It produces biased estimates of all three coefficients: $\hat{d}$ = 4.89 (s.e. = .45), $\widehat{b_1}$ = 12.41 (s.e. = .17), and $\widehat{b_2}$ = 7.48 (s.e. = .17). Note that the estimates are biased even though R^2 = .98 for the fitted model.

`tsls(friend ~ proximity + familiarity + similarity, ~ proximity + Z.fam + Z.sim)` uses instrumental variables to fit the regression of *friend* on *proximity*, *familiarity*, and *similarity* under the assumption that *proximity*, Z_{fam}, and Z_{sim} are exogenous variables. The instrumental-variables estimates are approximately accurate: $\hat{d}$ = 9.41 (s.e. = 1.11), $\widehat{b_1}$ = 10.48 (s.e. = .55), and $\widehat{b_2}$ = 5.11 (s.e. = .64).

Table A1

Random Sample of Psychology Articles Published in 2007 that Cite Baron and Kenny (1986)

		Independent	
Authors	Journal	variable	Mediator
Abele & Wojciszke	JPSP 93(5)	measured	measured
Birman & Taylor-Ritzler	Cul. Div. & Eth. Min. Psy. 13(4)	measured	measured
Bolger & Amarel	JPSP 92(3)	manipulated	measured and manipulated
Brendgen et al.	J. of Educ. Psy. 99(1)	measured	measured
Brockner et al.	J. of Applied Psy. 92(6)	measured	measured
Burns & Evon	Health Psy. 26(6)	measured	measured
Butler et al.	Emotion 7(1)	manipulated	measured
Caudle et al.	Am. J. Geriatr. Psychi. 15(8)	manipulated	measured
Clark & Kashima	JPSP 93(6)	manipulated	measured
Collins	J. of Applied Psy. 92(1)	measured	measured
Cullen & Hammer	J. of Occ. Health Psy. 12(3)	measured	measured
Downie et al.	Cul. Div. & Eth. Min. Psy. 13(3)	measured	measured
Flamenbaum & Holden	J. of Counseling Psy. 54(1)	measured	measured
Feeney	JPSP 92(2)	measured and manipulated	measured
Fu et al.	JPSP 92(2)	measured	measured

Authors	Journal	Independent variable	Mediator
Ganzel et al.	Emotion 7(2)	measured and manipulated	measured
Goldberg & Grandey	J. of Occ. Health Psy. 12(3)	manipulated	measured
Grundy et al.	J. of Family Psy. 21(4)	measured	measured
Hayes et al.	J. of Consul. & Clinic. Psy. 75(3)	measured	measured
Horowitz et al.	J. of Consul. & Clinic. Psy. 75(5)	manipulated	measured
Humphrey et al.	J. of Applied Psy. 92(5)	measured	measured
Kendall & Treadwell	J. of Consul. & Clinic. Psy. 75(3)	manipulated	measured
Lash et al.	Psy. Of Addictive Beh. 21(3)	manipulated	measured
Marigold et al.	JPSP 92(2)	manipulated	measured
Mendes et al.	JPSP 92(4)	measured	measured
Moody et al.	Emotion 7(2)	measured	measured
Plaks & Stecher	JPSP 93(4)	manipulated	measured
Pronin et al.	JPSP 92(4)	manipulated	measured
Ramrakha et al.	J. Am. Ac. Ch./Ado. Psychi. 46(10)	measured	measured
Rasmussen et al.	J. of Abnormal Psy. 116(4)	measured	measured
Reddy & Crowther	Cul. Div. & Eth. Min. Psy. 13(1)	measured	measured
Reisenzein & Studtmann	Emotion 7(3)	manipulated	measured
Rudman et al.	JPSP 93(5)	manipulated	measured
Blodgett Salafia et al.	J. of Counseling Psy. 54(4)	measured	measured

Authors	Journal	Independent variable	Mediator
Salzinger et al.	J. Am. Ac. Ch./Ado. Psychi. 46(7)	measured	measured
Sassenberg et al.	JPSP 92(2)	manipulated	measured
Schimel et al.	JPSP 92(5)	manipulated	measured
Sher et al.	J. of Abnormal Psy. 116(2)	manipulated	measured
Shin & Zhou	J. of Applied Psy. 92(6)	measured	measured
Smith et al.	J. Am. Ac. Ch./Ado. Psychi. 46(8)	manipulated	measured
Story et al.	Psychology and Aging 22(4)	measured	measured
Trim et al.	Psy. Of Addictive Beh. 21(1)	measured	measured
Twenge et al.	JPSP 92(1)	manipulated	measured
Vannoy et al.	Am. J. Geriatr. Psychi. 15(12)	manipulated	measured
Wedig & Nock	J. Am. Ac. Ch./Ado. Psychi. 46(9)	measured	measured
Zhang et al.	JPSP 92(1)	manipulated	measured