Journal of Personality and Social Psychology
2010, Vol. 98, No. 4, 550–558

0022-3514/10/$12.00 DOI: 10.1037/a0018933

Yes, But What's the Mechanism? (Don't Expect an Easy Answer)

John G. Bullock and Donald P. Green
Yale University

Shang E. Ha
Brooklyn College of the City University of New York

Psychologists increasingly recommend experimental analysis of mediation. This is a step in the right direction because mediation analyses based on nonexperimental data are likely to be biased and because experiments, in principle, provide a sound basis for causal inference. But even experiments cannot overcome certain threats to inference that arise chiefly or exclusively in the context of mediation analysis—threats that have received little attention in psychology. The authors describe 3 of these threats and suggest ways to improve the exposition and design of mediation tests. Their conclusion is that inference about mediators is far more difficult than previous research suggests and is best tackled by an experimental research program that is specifically designed to address the challenges of mediation analysis.

Keywords: mediation, experiments, causal inference, indirect effects

Supplemental materials: http://dx.doi.org/10.1037/a0018933.supp

A common criticism of experiments is that they reveal but do not explain causal relationships. Consider experimental demonstrations that social rejection causes aggressive behavior (DeWall, Twenge, Gitter, & Baumeister, 2009). Does the effect occur because rejection makes people angry, because it makes them more likely to perceive others' actions as hostile, or for other reasons? Such a question implies 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 (Blau & Duncan, 1967; Fisher, 1935; Lazarsfeld, 1955), but it is now more common in psychology than in any other discipline. The best-known article on the subject, by Baron and Kenny (1986), is the most frequently cited article in the history of the *Journal of Personality and Social Psychology*, and mediation analysis is now almost mandatory for new social-psychology manuscripts (Quiñones-Vidal, López-Garcia, Peñaranda-Ortega, & Tortosa-Gil, 2004).

Because of its prominence, mediation analysis has attracted more than the usual amount of scrutiny. The method advanced by Baron and Kenny (1986), although still widely used, has been subject to a host of criticisms that have led to refinements (e.g., Mathieu & Taylor, 2006; Maxwell & Cole, 2007; McDonald, 1997; Shrout & Bolger, 2002). Some critics, going further, champion experimental methods of mediation analysis on the ground that they permit stronger conclusions about mediation (Spencer, Zanna, & Fong, 2005; Stone-Romero & Rosopa, 2008; see also Aronson, Wilson, & Brewer, 1998, p. 105).

These are important developments, but they leave unresolved fundamental concerns about mediation analysis that increasingly animate statisticians (e.g., Robins, 2003; Rubin, 2005). For reasons explained below, applications of the Baron-Kenny method and other nonexperimental methods are likely to produce biased estimates of mediation effects. But even experiments cannot overcome certain threats to inference that arise only in the context of mediation analysis or that are especially problematic in that context. The accuracy of experimental mediation analysis depends on the ability of experimenters to manipulate one mediator without manipulating others. Even when experimenters succeed in targeting a particular mediator, their estimates of indirect effects typically apply to an unknown subset of subjects in their sample. And when the effects of treatments and mediators vary among subjects within a sample, even experimental designs can misstate the extent of mediation.

These problems are striking because they arise even in settings that are very favorable to mediation analysis: experiments in which both a treatment and a mediator are manipulated. Persistent threats to inference do not imply that mediation analysis is hopeless, but they do imply that impediments to understanding mediation are fundamental, rather than the consequences of particular statistical procedures or research designs. In practice, it is often impossible to draw conclusions about mediation without invoking strong and untestable assumptions. And even when these assumptions are invoked, the data requirements for persuasive mediation analysis typically entail drawing on numerous studies. Throughout this article, we therefore urge readers to think of mediation analysis as a cumulative enterprise. Persuasive conclusions about mediation are difficult to reach under any circumstances, but they are most likely to be reached when they derive from an experimental research program that addresses the particular challenges of mediation analysis—challenges that we describe here.

John G. Bullock and Donald P. Green, Department of Political Science, Yale University; Shang E. Ha, Department of Political Science, Brooklyn College of the City University of New York.

We thank Marilynn Brewer, Kelly Brownell, Geoff Cohen, Jack Dovidio, Ruth Ditlmann, Adam Glynn, Anthony Greenwald, Andrew Hayes, Lawrence James, Charles Judd, Alan Kazdin, David Kenny, Kevin King, David MacKinnon, Margaret McConnell, Ryan Moore, Mary Murphy, Elizabeth Levy Paluck, Patrick Rosopa, and Peter Salovey for their comments.

Correspondence concerning this article should be addressed to John G. Bullock or Donald P. Green, PO Box 208209, New Haven, CT 06520-8209. E-mail: john.bullock@yale.edu or donald.green@yale.edu

We proceed as follows. Using a minimal amount of formal notation, we explain why regression models applied to nonexperimental data are likely to produce biased estimates of mediation effects. We then consider experimental designs for mediation analysis. Although some limitations of these designs have been discussed before (e.g., Spencer et al., 2005; Kenny, 2008, p. 356), we focus on issues that have received little attention, drawing on recent developments in statistics. To minimize technical exposition, the statistical arguments that we advance are backed up by proofs that we present in the Supplemental Materials. We are not breaking new ground with respect to proofs and derivations. As we note later, all of the technical ideas that we present may be found elsewhere. Our aim is to show the practical relevance of these abstract statistical arguments and to offer recommendations for the design and presentation of mediation analysis.

Mediation Analyses With Unmanipulated Mediators Are Prone to Bias

Many multi-equation regression methods have been proposed to test whether one variable mediates the effect of another. (See MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002, for an overview.) Our criticism in this section applies to each of these "measurement-of-mediation" methods (Spencer et al., 2005, p. 846), but for simplicity of exposition, we focus on the well-known method proposed by Baron and Kenny (1986, p. 1177). Like many other measurement-of-mediation designs, it is based on three linear equations:

$$M_i = \alpha_1 + aX_i + e_{i1}, \tag{1}$$

$$Y_i = \alpha_2 + cX_i + e_{i2}, \text{ and} \tag{2}$$

$$Y_i = \alpha_3 + dX_i + bM_i + e_{i3}, \tag{3}$$

where i indexes subjects in the sample; Y_i is a dependent variable; X_i is an independent variable; M_i is a potential mediator of the independent variable; and α_1, α_2, and α_3 are intercepts. e_{i1}, e_{i2}, and e_{i3} are mean-zero error terms that represent the cumulative effect of omitted variables; for example, e_{i1} represents the effect on M_i of variables other than X_i. To sidestep the question of whether X truly causes Y, we assume throughout this article that X is randomly assigned.[1]

The coefficients of interest are a, b, c, and d. The effect of M_i on Y_i is b. The total effect of X_i on Y_i is c. The direct effect of X_i on Y_i is d. The indirect ("mediated") effect of X_i on Y_i is ab or, equivalently, $c - d$.

In the absence of sampling variability, the estimator of b used in applications of the Baron-Kenny procedure equals

$$b + \frac{\mathrm{cov}(e_1, e_3)}{\mathrm{var}(e_1)},$$

and the estimator of d equals

$$d - a\,\frac{\mathrm{cov}(e_1, e_3)}{\mathrm{var}(e_1)},$$

where $\mathrm{cov}(e_1, e_3)$ is the covariance of e_1 and e_3, and $\mathrm{var}(e_1)$ is the variance of e_1. The upshot is that the estimator of b is biased: Even in infinitely large samples, it equals the true value of b plus an additional quantity. The estimator of d is biased, too: It equals the true value of d minus an additional quantity. The estimators of b and d yield accurate results only when these additional quantities equal zero (i.e., only when e_1 and e_3 do not covary).[2]

In practice, e_1 and e_3 are almost certain to covary. If an unobserved variable affects both M and Y, it will be reflected in both error terms, causing those error terms to covary. Even if no unobserved variable affects both M and Y, the error terms may covary if M is merely correlated with an unobserved variable (e.g., another mediator). This warning has been issued before by those who write about mediation analysis (e.g., Judd & Kenny, 1981, p. 607; MacKinnon et al., 2002, p. 100), but it seems to have escaped the attention of the mainstream of the discipline. As Kenny (2008, p. 356) put it, many scholars "either do not realize that they are conducting causal analyses or they fail to justify the assumptions that they have made." (See also James, 2008; Muller, Judd, & Yzerbyt, 2005; Stone-Romero & Rosopa, 2008)[3]

Part of the problem seems to lie with a common intuition that, even if one is not experimentally manipulating a mediator, bias in mediation analysis is less likely if the independent variable is randomized. This is incorrect. Random assignment of X can ensure that X bears no systematic relationship to e_1 or e_3, but it says nothing about whether M or Y are systematically related to these error terms and thus nothing about whether mediation analyses are biased. This warning does not appear in Baron and Kenny (1986) or in most subsequent work on mediation. But it appears clearly in a rarely cited predecessor, 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).

Baron-Kenny estimates are prone not only to bias but to bias of the sort that *overstates* the extent of mediation. To see why, recall the bias term in the Baron-Kenny estimator of b is $\mathrm{cov}(e_1, e_3)/\mathrm{var}(e_1)$. If this term shares the sign of b—if it is positive where b is positive, or negative where b is negative—Baron-Kenny esti-

[1] Baron and Kenny (1986) do not index X, M, Y, or error terms by individuals (i), as we do here. This is a difference in notation, not substance. Subscripts will prove useful below, when we discuss the possibility that the true values of a and b vary across members of a sample.

[2] In keeping with standard practice, we assume throughout this article that applications of the Baron-Kenny procedure use ordinary least squares (i.e., linear regression) estimators of the parameters in Equations 1–3. Use of different estimators in nonexperimental mediation research will not cure the problem of bias but may change the bias from $\mathrm{cov}(e_1, e_3)/\mathrm{var}(e_1)$ to some other quantity.

[3] Stepping back from mediation analysis to the more general problem of regression, estimators tend to be biased when one controls for variables that are affected by the treatment, as users of the Baron-Kenny method do when they control for M in a regression of Y on X. This "post-treatment bias" is the subject of a well-developed literature in statistics (e.g., Rosenbaum, 1984), but it has largely escaped the attention of those who conduct mediation analysis. At root, it is one instance of an even more general rule: Regression estimates are unbiased only if the independent variables in the regression equations are independent of the error terms. 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).

mates of b are inflated. In practice, this quantity usually shares the sign of b, because factors other than X that affect M also affect Y in 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 education and income in the same direction: Proximity to good schools, parents' 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 tends 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 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.

The many improvements on the method of Baron and Kenny (1986) cannot protect researchers from making biased estimates of mediation effects when error terms covary, as is likely when X is randomly assigned but M is not. Techniques adapted for time-series and panel data (e.g., Maxwell & Cole, 2007) remain just as prone to bias due to unobservables that covary with the treatment. Bootstrapping and other small-sample methods (e.g., MacKinnon, Lockwood, & Williams, 2004) produce better standard errors for estimates that remain biased. Of course, some causal processes are more complex than the one described by Equations 1–3. For example, an independent variable may have multiple mediators, and the possibility of feedback in causal chains is often of interest. Recent decades have seen the development of structural equation modeling techniques to grapple with these complexities. These techniques have the virtue of modeling measurement error in sophisticated ways, but as several authors have pointed out (Luijben, 1991; Tomarken & Waller, 2005; Iacobucci, 2008), they do not resolve the problem of bias in mediation analysis caused by omitted variables. Nor do they address the substantial problems of mediation analysis that we describe in the next section.

Omitted-variables bias is the reason why scholars increasingly recommend experimental manipulation of mediators (e.g., Spencer et al., 2005; Stone-Romero & Rosopa, 2008; see also Aronson et al., 1998, p. 105). This is a step in the right direction: Experimental manipulation can ensure that mediators are uncorrelated with other variables, and in principle, it can generate unbiased estimates of direct and indirect effects. But it also invokes several subtle and potentially problematic assumptions, to which we now turn.

Practical and Conceptual Obstacles to Experimental Mediation Analysis

Nothing is new about the assertion that experiments make for better causal inference than observational studies. In Equation 3, for example, experimental manipulation of M permits unbiased estimation of b, the effect of M on Y. But to date, experimental recommendations have understated some conceptual difficulties of mediation analysis, three of which we consider here. First, analyses in which a mediator is experimentally manipulated will be inaccurate unless the experimental intervention affects only the mediator in question, and no other mediators. Second, experimental mediation analyses produce estimates of indirect effects that typically apply to only an unknown subset of subjects. Third, if subjects in a sample are differently affected by changes in X and M, even experiments that successfully manipulate a single mediator may produce inaccurate estimates of indirect causal effects. As we explain below, none of these difficulties can be framed as a simple matter of unobserved variables. The first two difficulties apply to all experimental analyses (rather than to just experimental mediation analyses), but they apply with special force to mediation analysis. And the third difficulty has no analog to problems that arise outside the study of mediation.[4]

To get a feel for the depth of these problems, consider the work of Bolger and Amarel (2007), who offer an elegant example of experimental mediation analysis. From previous research, they knew that social support often fails to reduce recipients' stress when it is "visible" (i.e., perceived as intended support; Bolger, Zuckerman, & Kessler, 2000). They hypothesized that the effect of visible support 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 generally does not increase; if it did, visible support might succeed in reducing stress. To manipulate subjects' feelings of efficacy, Bolger and Amarel had confederates speak to subjects who had been placed in a demanding achievement-related situation. (Subjects, all undergraduates, expected to give a speech that would be evaluated by an audience of graduate students.) They found that visible support was more likely to reduce stress when confederates spoke in an efficacy-promoting way and less likely to reduce stress when confederates spoke in an efficacy-diminishing way. The finding lends credence to their claim that efficacy mediates the effect of visible support. In general, the approach exemplified by Bolger and Amarel deserves much wider use. But its promise should be set against several important limitations.

First, experimental estimates of indirect effects are accurate only if the experimental interventions affect just the mediator in question and no other mediator. This requirement has been noted before in passing (e.g., Spencer et al., 2005, p. 847), but we wish to draw attention to several related points that have not been previously discussed in print.

Strong justification for the assumption that an experimental intervention is targeting a particular mediator can usually be had only when the intervention is specifically designed to affect the mediator, as with Bolger and Amarel's (2007) careful use of confederates' statements to promote or diminish their subjects' feelings of efficacy. But even this is no guarantee. For example, if confederates' statements also affect subjects' moods, and if mood also mediates the relation between visible support and stress reduction, the assumption will be violated, and estimates of the effect of efficacy will be contaminated by the effect of mood.

Of course, a version of this problem applies to all experimental analyses: If experimental interventions are to produce meaningful

[4] In their nuanced argument for experimental mediation analysis, Spencer et al. (2005) distinguish between two kinds of experimental studies of mediators. *Experimental-causal-chain* experiments are those in which a mediator is directly manipulated. *Moderation-of-process* experiments are those in which researchers manipulate a variable that in turn affects a mediator. Our arguments in this section (and, in the Supplemental Materials, our proofs) apply to both types of experiments.

causal inferences, they must target the particular factor whose effect we want to determine. But the problem is especially acute in mediation analysis, because mediators, are almost exclusively cognitive variables that cannot be directly observed. It is harder to determine whether an intervention has isolated just one variable when that variable and closely related variables cannot be directly observed. Note, too, that this problem cannot be reduced to a matter of omitted variables: An intervention either isolates the effect of a single mediator or it does not, and if it does not, observing new variables will do nothing to solve the problem.

The challenge posed by the presence of multiple mediators is fundamental for two reasons. First, although it is typically easy to believe that an independent variable's effect may be transmitted through multiple causal pathways, it is difficult to formulate a comprehensive model that includes all potential pathways. Second, even if one can describe in detail all of the causal pathways, measuring the elements of these pathways—that is, all of the potential mediators—is a daunting task. But this is what is required to ensure that estimates of mediation are accurate. Perhaps no one will fully succeed in these two tasks, but researchers should know that the extent to which they succeed at these tasks heavily affects the credibility of their analyses. We revisit this issue later.

Even if those who use the experimental approach succeed in targeting particular mediators, they confront a second dilemma: The approach produces estimates of indirect effects that apply not to the entire sample but only to those subjects who are affected by the experimental intervention (e.g., Angrist, Imbens, & Rubin, 1996). To see why, consider a clinical trial that we conduct to learn the effect of a pill. All treatment-group subjects are asked to take the pill. Some refuse. We have no way of learning how these "noncompliers" would have been affected by the pill if they had taken it. We are therefore unable to estimate the average effect of taking the pill for all subjects in our sample.[5]

This problem is less obvious but more acute in the case of mediation analysis. Just as we cannot learn how subjects would be affected by taking a pill if they do not take it, we cannot learn how—to return to Bolger and Amarel (2007)—subjects would be affected by a change in efficacy if we cannot change their level of efficacy. But in the clinical trial, we can at least know which subjects refuse to take the pill. The situation is more difficult in mediation analysis because most mediators are cognitive variables that cannot be directly observed, and noncompliance is therefore more difficult to observe, too. For example, we cannot know which subjects' feelings of efficacy remain unchanged by a confederate's statement—although this is, from the standpoint of causal inference, on par with refusing to take a pill. Thus, in addition to being unable to estimate an average indirect effect for all of our subjects, we often cannot even know *which* subjects our estimates apply to when we conduct a mediation analysis.

A related, unintuitive consequence of this predicament is that different experimental manipulations of M may produce different conclusions about mediation, even if the manipulations have the same average effect on M. This threat to inference arises whenever the different manipulations affect different groups of subjects within the sample. For example, if Bolger and Amarel (2007) had two scripts with which to increase the efficacy of their subjects, and the first script affected one subgroup of subjects, whereas the second script affected a different (perhaps overlapping) group of subjects, the scripts might well produce different conclusions about the mediating power of efficacy. And this would be so even if each script had the same average effect on efficacy.

This problem is acute because mediation analysis seeks to gauge the effect on Y of an X-induced change in M. We cannot simply manipulate X to gauge this effect: Recall that analysis based on X-only manipulation is subject to the biases that we detailed in the previous section. Instead, we must use another experimental intervention—call it Z—to induce change in M while holding X constant. The problem is that changes in X may affect M for one group of subjects, whereas changes in Z may affect M for a different group. Unless the relation between M and Y is the same for both groups, using Z to manipulate M will produce a misleading estimate of the way in which M mediates X.

Given these formidable difficulties, we are skeptical about the ability of isolated experiments—even experiments in which X and M are randomly assigned—to provide clear evidence of mediation. But if we seem unduly skeptical, note that we are only staking out a middle ground in arguments about the possibility of meaningful mediation analysis. The skeptical extreme is staked out by statisticians who increasingly raise a third objection that applies even when mediators are successfully manipulated and isolated. They note that Equations 1–3 have coefficients that do not vary from subject to subject (i.e., the coefficients in these equations are not indexed by i), implying that the effects of X on M and M on Y are equal for all members of the sample. There is typically no reason to make this assumption, just as there is typically no reason to assume that the effect of X on Y is the same for all subjects in a study. Ordinarily, when we are not examining mediation, the assumption of unvarying effects does little harm. Suppose that we conduct an experiment in which X is randomly manipulated and in which its effect on Y varies across subjects. If we regress Y on X, the coefficient on X simply indicates the average effect of X. On the other hand, if the effects of X and M vary across subjects, regression of Y on X and M will typically *not* recover the average direct effect of X, even if both X and M have been experimentally manipulated (see Glynn, 2009; Pearl, 2001; Robins, 2003).

To see why causal heterogeneity frustrates any attempt to calculate average direct and indirect effects, consider the models

$$M_i = \alpha_1 + a_i X_i + e_{i1}, \tag{1a}$$

$$Y_i = \alpha_3 + d_i X_i + b_i M_i + e_{i3}. \tag{3a}$$

These models are similar to Equations 1 and 3. The difference is that these models allow coefficients to vary from subject to subject, whereas Equations 1 and 3 do not. Given Equations 1a and 3a, the indirect effect of X on Y for any subject i is $a_i b_i$. We cannot observe a_i or b_i for any individual subject, but we can conduct one experiment that produces an estimate $\bar{a}$ of the average a_i, and another experiment that produces an estimate $\bar{b}$ of the average b_i. If a_i and b_i are the same for every subject, $\bar{a}\bar{b}$ will be an accurate

[5] Of course, we can estimate an average effect of assignment to the treatment group that applies to all subjects, but this is not the same as estimating the average effect of the pill. In practice, if we want to learn about the pill's effect, we are reduced to estimating the "local average treatment effect": the average effect of the pill among subjects who take the pill if they are assigned to the treatment group but not if they are assigned to the control group. A proof appears in Part 4 of the Supplemental Materials (see also Morgan & Winship, 2007).

estimate of the average direct effect (in the absence of sampling variability). But suppose that a_i and b_i are both negative for some subjects, both positive for others. In this case, there is a positive indirect effect of X for every subject in the sample, because $a_i b_i > 0$ for every subject. But $\bar{a}$ may be zero, negative, or positive. And $\bar{b}$, too, may be zero, negative, or positive. Consequently, the estimate of the average indirect effect ($\bar{a}\bar{b}$) may be zero or negative, even though the true indirect effect is positive not just on average but for every subject in the sample. Figure 1 illustrates the problem, and the SPSS code in Part 6 of the Supplemental Materials permits readers to explore how different parameter values change the sign and magnitude of the bias.

As Kenny, Korchmaros, and Bolger (2003, p. 118) and Bauer, Preacher, and Gil (2006, pp. 146–147) pointed out, the naive arithmetic of mediation analysis becomes problematic whenever there are different effects for X on M and M on Y. For example, Cohen (2003) wanted to understand how reference-group cues (X) affect attitudes toward social policy (Y). In his experiments, politically conservative subjects received information about a generous welfare policy. Some were told that the policy was endorsed by the Republican Party. Others received no endorsement information. Cohen's findings are consistent with cues promoting systematic elaboration (M) of the policy information and with systematic elaboration in turn promoting positive attitudes toward the policy (Cohen, 2003, p. 817). On the other hand, Mackie, Worth, and Asuncion (1990; see also Mackie, Gastardo-Conaco, & Skelly, 1992) and others suggested that reference-group cues inhibit systematic processing of information and that systematic processing promotes the influence of policy details, which should lead, in this case, to *decreased* approval of the generous welfare policy among the conservative subjects. For present purposes, there is no need to favor either of these theories or to attempt a reconciliation. We need only note that they suggest a case in which causal effects may be heterogeneous and in which the arithmetic of mediation accounting breaks down. Let some subjects in an experiment be "Cohens": For these people, exposure to reference-group cues heightens systematic processing (a_i is positive), and systematic processing makes attitudes toward a generous welfare policy more favorable (b_i is positive). But other subjects are "Mackies": For them, exposure to reference-group cues limits systematic processing (a_i is negative), and systematic processing makes attitudes toward a generous welfare policy less favorable (b_i is negative). Here again, the indirect effect is positive for every subject, because $a_i b_i > 0$ for all i. But if the experimental sample includes both Cohens and Mackies, $\bar{a}$ and $\bar{b}$ may each be positive, negative, or zero. The conventional estimate of the average indirect effect, $\bar{a}\bar{b}$, may therefore have the wrong sign.

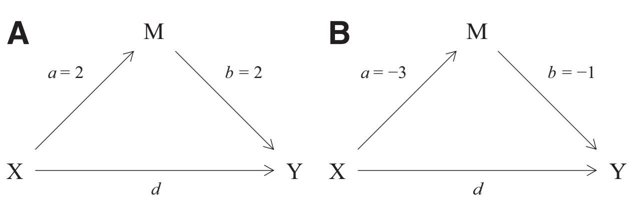

Figure 1. Even when the mediator is experimentally manipulated, causal heterogeneity makes it difficult to estimate indirect effects. For subjects represented by Panel A, $a = 2$ and $b = 2$. The indirect effect for these subjects is $ab = 4$. For subjects represented by Panel B, $a = –3$ and $b = –1$. The indirect effect for these subjects is $ab = 3$. The indirect effect is positive for every subject. But if there are equal numbers of Panel A and Panel B subjects, $\bar{a} = –.5$ and $\bar{b} = .5$. The conventional estimate of the average indirect effect is $\bar{a}\bar{b} = -.25$. This estimate has the wrong sign and is much smaller in magnitude than the true indirect effect, in spite of the randomization of both X and M. Nothing is special about the numbers used in this example; an infinite variety of other numbers would produce similar results.

Moreover, effects need not differ so sharply across members of a sample to make mediation analysis problematic. For example, consider a case in which the effects of X and M and M on Y are positive for all subjects. If these effects are negatively correlated—such that the first effect is large and the second effect small for some subjects, and vice versa for others—$\bar{a}\bar{b}$ may be large, whereas the true indirect effect for each subject may be small. This is another way in which the usual arithmetic of mediation analysis can go awry.

Discussions of "moderated mediation" (Muller et al., 2005) consider causal heterogeneity but seldom discuss the problem that it poses to the calculation of average mediation effects.[6] If X and M have been experimentally manipulated, and if their effects can be modeled as functions of observed variables and purely random error, and if one has a sufficient number of subjects (typically numbering well into the hundreds) for each level of sensitivity to changes in X and M, then methods of estimating moderated mediation can solve the problem posed by causal heterogeneity. But we know of no moderated-mediation studies that meet the first of these conditions, let alone all of them. The same is true of recent attempts to bring multilevel modeling to bear on the study of mediation (Bauer et al., 2006; Kenny et al., 2003; Krull & Mackinnon, 1999). Multilevel modeling treats variation in the effects of X on M and M on Y as random and strives to account for differences in these effects across subjects by using individual and group-level predictors. However, as Bauer et al. (2006, pp. 144–145, 158–59) and Kenny et al. (2003, p. 126) acknowledged, such models invoke very strong and often untestable modeling assumptions. For example, in addition to assuming that e_{i1} and e_{i3} are independent for each subject, multilevel modeling assumes that errors in regression equations are unassociated not just with predictors but with variation in coefficients as well (Bauer et al., 2006, p. 145).

None of these cautions implies that experiments are useless for mediation analysis. Indeed, we are convinced that the likely bias in measurement-of-mediation analyses warrants a further shift toward experiments. At the same time, one must bear in mind the formidable requirements that mediation analysis imposes on experimenters. We must devise manipulations that isolate specific mediators. This requires careful validation over multiple studies. Suppose we succeed: Our estimates of mediation will apply not to

[6] Few mediation analyses consider the possibility that treatment effects are heterogeneous. In a content analysis of 50 randomly selected social-psychology articles published in 2007, 46 contained some mediation analysis, but only three considered the possibility that mediation was moderated. Mediators were not experimentally manipulated in any of these three studies, leaving them subject to the concerns raised in the previous section. See Bullock, Green, and Ha (2008) for further details about the content analysis.

an entire sample but only to the subset of subjects who are affected by the manipulations. Finally, within-sample variation in the effects of X on M or M on Y may lead to biased mediation estimates. All of these problems (and more) apply in some form to nonexperimental analysis, too; we have focused on studies in which both X and M are experimentally manipulated because such studies are especially favorable to mediation analysis. The proper conclusion is that mediation is an *inherently* difficult subject—difficult even under favorable conditions and more difficult than the proliferation of regression-based and often-formulaic mediation analyses may suggest.

Recommendations

Some statistical procedures can help us to understand mediation, but the greatest obstacles to mediation analysis cannot be overcome solely or even chiefly by such procedures. Our recommendations therefore differ from those in recent textbooks and methodological discussions (e.g., Iacobucci, 2008; MacKinnon, 2008). In particular, we focus on how statistical *procedures* should be paired with statistical *arguments* and on how one might design a series of experiments to make a more convincing case for mediation:

1. Those who do not manipulate mediators should explain why $\text{cov}(e_1, e_3) = 0$—that is, why unobserved factors that affect M are uncorrelated with unobserved factors that affect Y. In addition, they should explain why the mediator that they observe (M) is uncorrelated with mediators that they have not observed. Such arguments are rarely made, but without them, it is difficult to maintain that nonexperimental mediation analysis generates credible results.

2. Those who experimentally manipulate mediators should explain why they believe that each manipulation is affecting only one mediator and not others. This entails describing the causal paths by which X may affect Y and explaining why each manipulation affects only one of these paths. The list of alternative causal paths may be extensive, and a succession of studies may be needed to demonstrate that a given intervention tends not to affect these alternative paths.

3. Experimental mediation analysis entails changing the value of M while holding X constant. This is problematic because the theoretically relevant changes in M are induced by changes in X. Experimental mediation analysis requires us to find another way of changing the values of M, but there is no guarantee that these changes will be the same as those that would be induced by changes in X. Changes in X may affect the value of M for one group of subjects, whereas the experimenter's manipulation of M—call it Z—may affect the value of M for a different group of subjects. In this case, even experimental mediation analyses may produce misleading inferences about indirect effects.

   Rarely, if ever, can one directly test the assumption that changes in X and changes in Z affect M for the same group of subjects. But one can address the issue by conducting studies that examine differently induced changes in M, and we recommend greater emphasis on this aspect of experimental design. For example, we can ask whether an additional year of schooling has the same average effect on achievement when that year is induced by a law requiring attendance, by giving money to families whose children stay in school, or by a preschool program that enhances school readiness (e.g., Krueger & Whitmore, 2001; Miguel & Kremer, 2004). If an additional year has the same average effect in each setting, it becomes more plausible to think that all ways of changing M affect the same group of subjects. On the other hand, if one finds variation across interventions and contexts, one must be more cautious about drawing inferences about the effects of M, as they seem contingent on the way in which changes in M are induced.

4. Those who analyze mediation should recognize that if the effects of X and M vary from subject to subject within a sample, it may be misleading to estimate the average direct or indirect effects for the entire sample. To determine whether heterogeneous effects are a problem, we recommend examining the effects of X and M among different groups of subjects. If these effects differ little from group to group (e.g., from women to men, authoritarians to nonauthoritarians), we become more confident that causal heterogeneity is not affecting our analysis. On the other hand, if there are large between-group differences in the effects of X and M, estimates of the average indirect effect may be inaccurate even if they are derived from an experiment in which both X and M are manipulated. One might use multilevel modeling to account for between-group differences, but we are reluctant to embrace the strong assumptions that multilevel modeling invokes (e.g., Bauer et al., 2006, pp. 144–145, 158–159; Kenny et al., 2003, p. 126). We recommend instead that researchers try to identify relatively homogeneous subgroups and make inferences about indirect effects for each subgroup rather than a single inference about an average indirect effect for an entire sample. Partitioning one's sample in this way reduces the power of statistical analyses, thereby making replication of findings and the pursuit of an integrated research program all the more important.

5. In principle, the problem posed by heterogeneity of the effects of X on M and M on Y can be solved by studies in which subjects are exposed over time to different, experimentally determined values of X and M. (See, e.g., Ratkowsky, Evans, & Alldredge, 1993.)

   Such designs can overcome the heterogeneity problem because they permit measurement of an indirect effect for each subject, which is not possible when subjects are exposed to X or M only once. But repeated-measures experimental designs are difficult to execute in many domains of social psychology. Even if they can be executed, their use entails a set of assumptions that may be difficult to meet, including the assumptions that within-person effects do not change over time and that the effect of each manipulation of X and M has "worn off" before the next manipulation is administered. Nevertheless, designs that expose subjects to repeated interventions are underused in the social sciences, and they have the potential to speak to the question of heterogeneous treatment effects. Our broader recommendation is that researchers be on the lookout for creative experimental designs that can speak persuasively to the question of heterogeneous treatment effects.

Taken together, the challenges of experimental mediation analysis—crafting interventions that isolate particular mediators, determining which subjects in a sample are affected by the interventions, and accounting for the possibility that causal effects covary within a sample—are formidable. That said, they are already met to varying extents by a number of literatures. Research on stereotype threat is an especially promising example. In under two decades, this literature has identified a main effect. It has shown that the effect can be induced by different experimental interventions (e.g., by varying information about the gender neutrality of a test or by changing the gender composition of fellow test takers). The presence of an effect under different interventions suggests that causal heterogeneity may not be a grave problem. Moreover, the literature has also identified many potential mediators, including anxiety (Spencer, Steele, & Quinn, 1999), working memory (Schmader & Johns, 2003), dejection (Keller & Dauenheimer, 2003), and arousal (Ben-Zeev, Fein, & Inzlicht, 2005; Blascovich, Spencer, Quinn, & Steele, 2001; O'Brien & Crandall, 2003). And some of these variables, too, have effects that can be induced via different interventions. For example, we now know that many different interventions can induce arousal.

What remains is to integrate these findings in ways that give us better purchase on potential mediators. Doing so entails more than amassing a heap of studies about stereotype threat. It entails a research agenda that targets the particular challenges of mediation analysis. Specifically, it entails crafting interventions that affect one likely mediator—say, arousal—while leaving the others unaffected. It entails examining the average effects of these interventions on different subgroups to see whether the interventions—and thus mediation analyses themselves—apply to entire samples or just to within-sample subgroups. It also entails further examining affected subgroups to see whether they are affected to the same or to different degrees by the variables of theoretical interest. If the latter, the aim of mediation analysis in stereotype-threat research should be to estimate indirect effects separately for different subgroups rather than to estimate average indirect effects for entire samples. These are demanding tasks, but they are feasible given the rapid growth of research on stereotype threat.

Experimental research programs that are integrated in this way will help us to learn about mediation; less obviously, they also stand to teach us about nonexperimental mediation analysis. They can do so by producing benchmark estimates of mediation effects against which we can gauge the accuracy of nonexperimental estimates. Experimental research has already produced thousands of benchmark estimates of the overall effects of independent variables, and these benchmarks have been used to render judgments about the accuracy (or lack thereof) of nonexperimental methods (e.g., Arceneaux, Gerber, & Green, 2006; Green, Leong, Kern, Gerber, & Larimer, 2009; LaLonde, 1986). There are no analogous benchmarks in mediation research because experimental mediation analysis is rare and because the systematic, cumulative use of experiments to study mediation has not yet been undertaken. The production of such benchmarks awaits the integrated use of multiple experiments to speak to the inherent difficulties of mediation analysis.

Discussion

In the decades since the publication of Baron and Kenny (1986), most mediation analysis has been guided by a multi-equation regression framework. This framework imposes assumptions that are unlikely to be satisfied in most psychological applications. Typically, mediating variables are correlated with unobserved variables that affect outcomes, in which case this approach misstates the mediator's role in the causal process.

Recognizing this limitation, scholars increasingly recommend experimental manipulation of mediators (e.g., Spencer et al., 2005; Stone-Romero & Rosopa, 2008; see also Aronson et al., 1998, p. 105). The experimental approach to mediation can overcome the unobserved-variables problem, and for that reason, it is in principle an improvement over the measurement-of-mediation approach. But it relies on assumptions that seem to be insufficiently appreciated. First, experimental manipulations that are used in mediation analysis must affect one mediator without affecting others. Second, if an experimental intervention affects only some members of a sample, estimates of indirect effects apply to only those members, rather than to the entire sample. Third, one cannot estimate an average indirect effect for an entire sample if the effects of treatments and mediators vary among members of the sample. When these problems arise, the usual rules for interpreting direct and indirect effects break down.

The general point is not that experiments are uninformative or infeasible. It is that mediation analysis is challenging because of difficulties that are fundamental to the endeavor rather than limitations of a particular procedure. All of the limitations that we have discussed are exacerbated in nonexperimental research. But precisely because experimentation simplifies analysis, it is the proper framework for highlighting complications that are more acute in mediation analysis than in other types of causal analysis.

These complications cannot be overcome by purely statistical innovations, and our recommendations therefore encourage awareness of important assumptions and their implications for research design. Those who conduct nonexperimental mediation analyses should explicitly justify the assumption that the mediator under examination is uncorrelated with other variables that may affect the outcome. Those who conduct experimental analyses should explicitly justify the assumption that their manipulations affect only one mediator. Experimenters should know that different manipulations of M may lead to different conclusions about mediation (even if these manipulations have the same average effect on M); they can therefore strengthen their conclusions by manipulating M in multiple ways. Finally, different reactions to independent variables and mediators among subjects within a sample can make it impossible to calculate a meaningful average indirect effect for all members of the sample. If subjects are believed to vary in their reactions to independent variables and mediators, researchers should estimate average indirect effects for homogeneous subgroups rather than a single average for the entire sample. None of these recommendations is easy to implement, but collectively, they can make mediation analysis more persuasive by leading researchers to speak directly to the assumptions on which mediation analysis depends.

In closing, we note that although none of these arguments is novel among statisticians, social scientists are largely unaware of them. Recent decades have seen dramatic growth of interest in mediation and some advance in the degree of technical sophistication that researchers bring to mediation analysis. Yet the quality of argumentation remains inadequate because researchers have not come to grips with some of the key assumptions on which their analyses depend. Deficient argumentation in turn leads to insufficient attention to issues of design. Assessing mediation is a conceptually deep and empirically vexing task, and those impatient for answers to questions of mediation seem to underestimate the challenges presented by the study of causal pathways.

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,* 444–455.

Arceneaux, K., Gerber, A. S., & Green, D. P. (2006). Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis, 14,* 37–62.

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* (4th ed., Vol. 1, pp. 99–142). New York, NY: 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,* 1173–1182.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11,* 142–163.

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41,* 174–181.

Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. M. (2001). African Americans and high blood pressure: The role of stereotype threat. *Psychological Science, 12,* 225–229.

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

Bolger, N., & Amarel, D. (2007). Effects of social support visibility on adjustment to stress: Experimental evidence. *Journal of Personality and Social Psychology, 92,* 458–475.

Bolger, N., Zuckerman, A., & Kessler, R. C. (2000). Invisible support and adjustment to stress. *Journal of Personality and Social Psychology, 79,* 953–961.

Bullock, J. G., Green, D. P., & Ha, S. E. (2008). *Experimental approaches to mediation: A new guide for assessing causal pathways.* Unpublished manuscript, Yale University, New Haven, CT.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology, 85,* 808–822.

DeWall, C. N., Twenge, J. M., Gitter, S. A., & Baumeister, R. F. (2009). It's the thought that counts: The role of hostile cognition in shaping aggressive responses to social exclusion. *Journal of Personality and Social Psychology, 96,* 45–59.

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

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

Glynn, A. N. (2009). *The product and difference fallacies for indirect effects.* Unpublished manuscript, Harvard University, Cambridge, MA.

Green, D. P., Leong, T. Y., Kern, H. L., Gerber, A. S., & Larimer, C. W. (2009). Testing the accuracy of regression discontinuity analysis using experimental benchmarks. *Political Analysis, 17,* 400–417.

Iacobucci, D. (2008). *Mediation analysis.* New York, NY: Sage.

James, L. R. (2008). On the path to mediation. *Organizational Research Methods, 11,* 359–363.

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

Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat on women's math performance. *Personality and Social Psychology Bulletin, 29,* 371–381.

Kenny, D. A. (2008). Reflections on mediation. *Organizational Research Methods, 11,* 353–358.

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods, 8,* 115–128.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. *Economic Journal, 111,* 1–28.

Krull, J. L., & Mackinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review, 23,* 418–444.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76,* 604–620.

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. 115–125). Glencoe, IL: Free Press.

Luijben, T. C. W. (1991). Equivalent models in covariance structure analysis. *Psychometrika, 56,* 653–665.

Mackie, D. M., Gastardo-Conaco, M. C., & Skelly, J. J. (1992). Knowledge of the advocated position and the processing of in-group and out-group persuasive messages. *Personality and Social Psychology Bulletin, 58,* 145–151.

Mackie, D. M., Worth, L. T., & Asuncion, A. G. (1990). Processing of persuasive in-group messages. *Journal of Personality and Social Psychology, 58,* 812–822.

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

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.

Mathieu, J. E., & Taylor, S. R. (2006). Clarifying conditions and decision points for mediational type inferences in organizational behavior. *Journal of Organizational Behavior, 27,* 1031–1056.

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.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica, 72,* 159–217.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference.* New York, NY: Cambridge University Press.

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.

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal:

Effects on women's math performance. *Personality and Social Psychology Bulletin, 29,* 782–789.

Pearl, J. (2001). Direct and indirect effects. In J. S. Breese & D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufman.

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,* 435–452.

Ratkowsky, D. A., Evans, M. A., & Alldredge, J. R. (1993). *Cross-over experiments: Design, analysis, and application.* New York, NY: Marcel Dekker.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. J. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 70–81). New York, NY: Oxford University Press.

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,* 656–666.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85,* 440–452.

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

Spencer, S. J., Steele, C. M., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4–28.

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,* 326–352.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and weaknesses. *Annual Review of Clinical Psychology, 1,* 31–65.

Received May 2, 2009
Revision received December 9, 2009
Accepted December 10, 2009 ■