Enough Already about "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose

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

The ANNALS of the American Academy of Political and Social Science 2010; 628; 200
DOI: 10.1177/0002716209351526

Enough Already about "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose

By
DONALD P. GREEN,
SHANG E. HA,
and
JOHN G. BULLOCK

The question of how causal effects are transmitted is fascinating and inevitably arises whenever experiments are presented. Social scientists cannot be faulted for taking a lively interest in "mediation," the process by which causal influences are transmitted. However, social scientists frequently underestimate the difficulty of establishing causal pathways in a rigorous empirical manner. We argue that the statistical methods currently used to study mediation are flawed and that even sophisticated experimental designs cannot speak to questions of mediation without the aid of strong assumptions. The study of mediation is more demanding than most social scientists suppose and requires not one experimental study but rather an extensive program of experimental research.

Keywords: mediation; causal mechanisms; causal inference; experiments

An academic phenomenon that occurs with astonishing regularity may be observed whenever experimental researchers present evidence of a causal effect. Inevitably, someone in the audience asks what mediating factor or factors explain this effect. The stronger the experimental effect, the greater the audience's interest in mediators. When confronted with experimenters who fail to offer evidence that explains how their intervention's effect is transmitted, audiences routinely grumble about "black box" experimentation.

One can scarcely fault scholars for expressing curiosity about the mechanisms by which an experimental treatment transmits its influence. After all, many of the most interesting discoveries in science have to do with identifying mediating factors in a causal chain. For example, the introduction of limes into the diet of seafarers in the eighteenth century dramatically reduced the incidence of scurvy, and eventually twentieth-century scientists figured out that the key mediating ingredient was vitamin C. Equipped with knowledge about why an experimental treatment works, scientists may devise

DOI: 10.1177/0002716209351526

other, possibly more efficient ways of achieving the same effect. Modern seafarers can prevent scurvy with limes or simply with vitamin C tablets.

Arresting examples of mediators abound in the physical and life sciences. Indeed, not only do scientists know that vitamin C mediates the causal relationship between limes and scurvy, they also understand the biochemical process by which vitamin C counteracts the onset of scurvy. In other words, mediators themselves have mediators. Physical and life scientists continually seek to pinpoint ever more specific explanatory agents.

Social scientists, too, are eager to pinpoint causal mechanisms; but, unfortunately, well-established claims about mediation remain relatively rare in the social sciences. We use the phrase "well-established" to denote claims that are backed up with compelling scientific evidence, not just claims that are widely believed because they appeal to widely held presuppositions. The notion that there is a dearth of compelling examples of mediation is doubtless a minority viewpoint. As Bullock and Ha (forthcoming) point out in their recent review of the mediation literature in political science, confident claims about this or that mediating variable abound in social science journal articles and literature reviews. Recent years have in fact seen growing enthusiasm for regression models that purport to establish claims about mediation; Malhotra and Krosnick (2007), for example, argue forcefully that this form of regression analysis ought to become more prominent in studies of electoral politics.

Despite their growing popularity, these regression models rest on naïve assumptions. The point of this article is to puncture the widely held view that it is a relatively simple matter to establish the mechanism by which causality is transmitted. This means puncturing the faith that has been placed in commonly used statistical methods of establishing mediation.

Fortunately, the algebraic groundwork for our argument may be found in the statistical literature on mediation that is largely unknown to social scientists. (See,

---

Donald P. Green is the A. Whitney Griswold Professor of Political Science at Yale University. Since 1996, he has served as director of Yale's Institution for Social and Policy Studies, an interdisciplinary research center that emphasizes field experimentation. In collaboration with his Yale colleagues, he has conducted an array of field experiments in political science, communication, criminology, and education.

Shang E. Ha received his Ph.D. from the University of Chicago in 2007. He is an assistant professor of political science at Brooklyn College–City University of New York. In 2007-09, he was a postdoctoral associate at the Institution for Social and Policy Studies at Yale University. His research focuses on voting behavior and political psychology, racial attitudes, and experimental methods. His work has appeared in the American Political Science Review, American Politics Research, *and* Political Research Quarterly.

John G. Bullock is an assistant professor of political science at Yale University and a Resident Fellow of Yale's Institution for Social and Policy Studies. He studies political psychology and public opinion, focusing on the ways in which new information affects people's political views. Much of his current research considers circumstances under which information can cause people's views to polarize. Related lines of inquiry use experiments to examine the public's level of political knowledge and the effects of political misinformation.

e.g., Holland 1988; Jo 2008; Sobel 2008; for a less technical overview of key issues, see Bullock, Green, and Ha 2009.) We will largely dispense with equations and try to state intuitively what others have stated formally. Our aim is to convince the reader of three things:

1. Conventional regression approaches to the study of mediation rely on strong and often implausible assumptions, even when applied to data in which a causal factor has been manipulated experimentally.
2. The natural progression of an experimental agenda makes it impractical to examine mediators until a causal relationship is firmly established.
3. Even when causal relationships are firmly established, demonstrating the mediating pathways is far more difficult—practically and conceptually—than is usually supposed.

Our argument is not that the search for mediators is pointless or impossible. Establishing the mediating pathways by which an effect is transmitted can be of enormous theoretical and practical value, as the vitamin C example illustrates. Rather, we take issue with the impatience that social scientists often express with experimental studies that fail to explain why an effect obtains. As one begins to appreciate the complexity of mediation analysis, it becomes apparent why the experimental investigation of mediators is slow work. Just as it took more than a century to discover why limes cure scurvy, it may take decades to figure out the mechanisms that account for the causal relationships observed in social science.

Conventional Approaches to the Study of Mediation Are Prone to Bias

Perhaps the most startling fact about the statistical investigation of mediation in the social sciences is how frequently it is attempted. Although path analysis goes back several decades, mediation analyses surged in popularity in the 1980s with the publication of Baron and Kenny (1986), which now ranks as the most frequently cited article ever to appear in the *Journal of Personality and Social Psychology*. The framework described by Baron and Kenny involves a series of regressions. First, one regresses the outcome (Y) on the independent variable (X). Upon finding an effect to be explained, one proposes a possible mediating variable (M) and regresses it on X. If X appears to cause M, the final step is to examine whether the effect of X becomes negligible when Y is regressed on both M and X. If M predicts Y and X does not, the implication is that X transmits its influence through M.

This type of analysis rests on a number of strong assumptions. The most contentious assumption is the requirement that M be independent of unmeasured factors that affect Y. Let's consider what this assumption means in practice for an experiment in which X is manipulated randomly. (Applying this analysis to observational data in which X is not randomized jeopardizes the premise of the investigation of mediators, namely, that X in fact exerts a *causal* influence on Y. We

consider experimental applications, as these in theory have a better chance of success.) Suppose one were interested in explaining why voter mobilization activity (X) affects electoral participation (Y), and imagine that mobilization activity were varied randomly so that there is no concern about whether the causal relationship between X and Y is real. One could posit a mediating pathway whereby get-out-the-vote campaigns enhance interest in political affairs (M), which in turn increases one's propensity to vote.

The problem with establishing this claim empirically is that there may be other mediators, each of which is correlated with interest in politics. For this example, a short list of additional mediators might include cognitive skills, feelings of internal efficacy, social ties to people who are politically engaged, and so forth. Unless one measures and controls for each of these alternative mediators, one risks attributing to political interest mediating effects that in fact flow through some other intervening factor. Of course, as a practical matter, it is impossible to measure all of the possibly confounding mediating variables. Putting measurement aside, it is rare that a researcher will be able to *think* of all of the confounding mediators.

When applied to data in which M is observed but not manipulated randomly, this kind of mediation analysis amounts to an ill-defined procedure with no clear stopping rule or method for detecting bias. Uncertain about the causal pathways and perhaps even the direction of causality, researchers tend to consider a variety of mediators, sometimes one at a time or in different combinations. From these analyses, a conclusion emerges about the successfulness with which one or more mediators explain the bivariate relationship between X and Y.

This type of analysis is vulnerable to two important critiques. The first concerns omitted variables. If M is positively correlated with unobserved causes of Y, its effect on Y may be exaggerated while the effect of X on Y is underestimated. That pattern of biases will tend to make the mediation analysis look more successful than it really is. This kind of bias seems to be quite common, for one of the ways in which researchers look for mediators is to consider variables that are correlated with Y. One reason that M may be correlated with Y is that they both are correlated with unobserved confounders.

A second line of critique is that M is poorly measured, which may lead to an underestimate of M's effect and the mistaken conclusion that factors other than M account for the relationship between X and Y. When several of the mediators are correlated and mismeasured, biases can be unpredictable in sign and magnitude. The use of structural equation modeling with latent variables is often hailed as a way to address the dubious assumptions underlying mediation analysis. Structural equation modeling is a step in the right direction insofar as it addresses the problem of measurement error, but structural equation models typically do nothing to address the problem of omitted variables.

Given the strong requirements in terms of model specification and measurement, the enterprise of "opening the black box" or "exploring causal pathways" using endogenous mediators is little more than a rhetorical exercise. We are at a loss to produce even a single example in political science in which this kind of

mediation analysis has convincingly demonstrated how a causal effect is transmitted from *X* to *Y*. What we have instead is a long list of examples in which mediation is proved with the aid of very strong and untested assumptions. The question is whether the situation improves as we move from observational designs to experimental designs, where both *X* and *M* are manipulated randomly.

Experimental Studies of Mediation Are Difficult to Design and Execute

In principle, experiments are the gold standard for estimating causal parameters, and so one naturally turns to experiments to assess hypotheses about mediation. Experiments can play a useful role in the study of mediators. If one is interested in whether *M* mediates the effects of *X* on *Y*, it makes sense to randomly manipulate *M* in order to see whether it indeed affects *Y*. It also seems sensible to manipulate *X* in order to gauge whether *Y* changes as a result. If one is prepared to assume that the causal effects of *X* and *M* are the same across subjects, this kind of "double experiment" can be quite informative. Finding that *M* affects *Y* suggests that *M* may be among the mediators of *X*. And finding that *X* affects *M* further suggests that *M* may transmit *X*'s influence to *Y*.

Two complications make this type of experimental investigation challenging. First, as a practical matter, it is seldom easy to design an experiment to manipulate *M*. Or, to put it more precisely, it is seldom easy to design an experiment that manipulates only *M* and not some other *M*' that might also mediate the effect of *X*. To return to the mobilization and voting example, suppose a researcher sought to assess the mediating effects of political interest. Producing an increase in political interest is no mean feat, and the task becomes especially challenging if one strives to generate interest in politics without inadvertently producing a change in political efficacy or political knowledge or any of the attitudinal correlates that might also mediate the effects of campaign contact. Note that this problem of experimental design is analogous to an identification problem in a simultaneous equations system. The more mediators one seeks to assess, the more elaborate one's experimental design must be, with multiple interventions designed to influence different mediators to different degrees.

Recent textbooks that discuss mediation too often skip over the problem of multiple mediators or send mixed messages about the difficulty of manipulating and measuring the mediators. For example, MacKinnon (2008) notes in passing (p. 66) that mediation models are sensitive to omitted variables bias but devotes his analysis of single- and multiple-mediator systems to the technical questions of how to compute the estimates and their standard errors. His proof of the unbiasedness of the regression approach (pp. 86-89) blithely assumes that *M* is unrelated to unmeasured causes of *Y*. Threats to inference are revisited more than 250 pages later, after MacKinnon explicates a number of techniques and empirical examples that presuppose strong modeling assumptions.

A second and related complication arises when we relax the assumption that the causal effects of *X* and *M* are the same for all subjects. The key idea is that we can identify only the causal effect of the *randomly induced* variation in these variables. Suppose, for example, that we were to induce heightened political interest by means of a multimedia presentation about the important ways in which elections shape political outcomes. Some subjects might become more politically interested, but others might be unaffected. Whatever downstream consequences of political interest on political participation we observe would reflect the change in behavior among those who were affected by the presentation. The subset of the population that is moved by one presentation might not be the same subset that is moved by another, and different subgroups might transmit their newfound political interest into participation in different ways. Tempting as it is to draw broad conclusions about mediation based on a single intervention, those conclusions really hold for only a subset of the population. The possibility of different treatment effects for different subgroups is in principle an empirical question. With enough experimental interventions, a researcher can gauge the extent to which the effects of *M* on *Y* or *X* on *M* or *X* on *Y* vary according to the way in which *X* and *M* are manipulated, but conducting an array of experiments is a formidable undertaking. It is a far cry from the run-some-regressions advice that comes from those in the thrall of the Baron-Kenny method.

For Those Still Not Discouraged, It Gets Worse

The previous section raised the nettlesome possibility that treatment effects may differ across subjects. Scholars working on mediation sometimes call this phenomenon "moderated" mediation in the sense that the causal paths vary in strength across subjects (Muller, Judd, and Yzerbyt 2005). Usually, however, when empirical researchers talk about moderated mediation, they are quick to assume that variation in effect size can be modeled as a function of observable factors. There is nothing wrong with trying to model interactions between measured variables, but the problem of unobserved variables remains. Unobserved sources of variation in effect size can throw off any attempt to draw inferences about mediation.

Consider the following example. Imagine we have a large sample, say, 10,000 observations. Suppose that for the odd-numbered observations, the data generation process looks as follows:

$$Y = M + u, \tag{1}$$

$$M = X + e. \tag{2}$$

In other words, we have set up the example such that a one-unit change in *M* leads to a one-unit change in *Y*. And a one-unit change in *X* leads to a one-unit change in *M*. In this example, *M* fully mediates the effect of *X* on *Y*. The model

contains two unobserved disturbance terms, u and e. In the spirit of making this example mimic an ideal experiment, suppose these two unobserved factors were independent of one another and drawn independently for each observation.

The data generation process for the even-numbered observations is similar, with one twist. The data generation process for u and e is the same as above; they are independent of one another and across observations. This time, however, the slopes are different:

$$Y = -M + u, \tag{3}$$

$$M = -X + e. \tag{4}$$

The total effect of X on Y is 1.0—the product of its negative effect on M and M's negative effect on Y. M fully mediates the effect of X.

What happens when we analyze all 10,000 observations without regard to the fact that half of the data are generated by the "odd" model and half by the "even" model? In short, we get misleading results. The total effect of X on Y is found to be 1.0, suggesting that there is a relationship in need of explanation. However, the regression of M on X suggests that X has no effect on M. And the regression of Y on both M and X indicates that X has an effect of 1.0 while M has an effect of 0. The implication is that M plays no role in transmitting X's influence to Y, but we know from the model that this is false.

This doubtless seems like an extreme example. In practice, we would not expect an unobserved factor to partition our sample in half, such that each half is subject to equal and opposite parameters. What is troubling about this example, however, is that one can come up with a range of different results simply by varying the proportion of people in the sample who are subject to each of the data generation processes. For example, if one-fifth of the sample is generated by equations (1) and (2) and four-fifths is generated by equations (3) and (4), a Baron-Kenny analysis will indicate that approximately half of X's influence remains unmediated, which is still incorrect.

The bottom line is that when subjects are governed by different causal laws, analyses that presuppose that the same parameters apply to all observations may yield biased results. Experimental design is helpful insofar as it helps avoid some of the most common sources of bias, such as correlation between M and u. But a single experiment is unlikely to settle the question of heterogeneous treatment effects. In order to ascertain whether different subjects transmit the causal influence of X in different ways, multiple experiments—maybe decades' worth—will be necessary.

Conclusion

Experimenters have good reason to be cautious when encouraged to divert attention and resources to the investigation of causal mechanisms. First, black

box experimentation as it currently stands has a lot going for it. One can learn a great deal of theoretical and practical value simply by manipulating variables and gauging their effects on outcomes, regardless of the causal pathways by which these effects are transmitted. Introducing limes into the diet of seafarers was an enormous breakthrough even if no one at the time had the vaguest understanding of vitamins or cell biology. Social science would be far more advanced than it is today if researchers had a wealth of experimental evidence showing the efficacy of various educational, political, or economic interventions—even if uncertainty remained about why these interventions work.

Second, the rush to study mechanisms presupposes that experiments have to date established these basic causal relationships in need of explanation. This is far from the case, even in relatively well-developed experimental subfields. Critics of "mere" black box experimentation fail to realize that nailing down an experimental effect with precision takes a great deal of sustained effort. For any researcher working in the early phases of an experimental research program, devoting resources to the manipulation of mediators (and investigation of subgroup differences in causal effects) is a gamble, as there is no guarantee that the experimental intervention will produce a substantively interesting average effect on the outcome. Few experimental programs in social science are sufficiently advanced to warrant this kind of gamble.

A more judicious approach at this juncture in the development of social science would be to encourage researchers to measure as many outcomes as possible when conducting experiments. For example, consider the many studies that have sought to increase voter turnout by means of some form of campaign contact, such as door-to-door canvassing. In addition to assessing whether the intervention increases turnout, one might also conduct a survey of random samples of the treatment and control groups in order to ascertain whether these groups differ in terms of interest in politics, feelings of civic responsibility, knowledge about where and how to vote, and so forth. With many mediators and only one intervention, this kind of experiment cannot identify which of the many causal pathways transmit the effect of the treatment, but if certain pathways are unaffected by the treatment, one may begin to argue that they do not explain why mobilization works. As noted above, this kind of analysis makes some important assumptions about homogeneous treatment effects, but the point is that this type of exploratory investigation may provide some useful clues to guide further experimental investigation.

As researchers gradually develop intuitions about the conditions under which effects are larger or smaller, they may begin to experiment with variations in the treatment in an effort to isolate the aspects of the intervention that produce the effect. For example, after a series of pilot studies that suggested that social surveillance might be effective in increasing voter turnout, Gerber, Green, and Larimer (2008) launched a study in which subjects were presented one of several interventions. One encouraged voting as a matter of civic duty; another indicated that researchers would be monitoring who voted; a third revealed the voting behavior of all the people living at the same address; and a final treatment revealed the

voting behavior of those living on the block. This study stopped short of measuring mediators such as one's commitment to norms of civic participation or one's desire to maintain a reputation of an engaged citizen; nevertheless, the treatments were designed to activate mediators to varying degrees. One can easily imagine variations in this experimental design that would enable the researcher to differentiate more finely between mediators. And one can imagine introducing survey measures to check whether these inducements produce an intervening psychological effect consistent with the posited mediator.

So long as the limitations of this exploratory mode of investigation are clear, scientific investigation can proceed in an orderly manner. The problem is that so long as social scientists operate with a mistaken understanding of what can be expected from a mediation analysis, they will flit from one topic to another without an appropriate sense of the limits of what has been learned along the way. When critics make pious declarations about the importance of opening the black box, one must recognize that in social sciences black boxes are rarely if ever opened. Sometimes they are *declared* open by researchers who are too sanguine about the power of their lock-picking skills. Such declarations give the impression that the work is easy or already complete, which ironically slows the painstaking process by which real progress is made.

References

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

Bullock, John G., Donald P. Green, and Shang E. Ha. 2009. Experimental approaches to mediation: A new guide for assessing causal pathways. Unpublished manuscript, Yale University, New Haven, CT.

Bullock, John G., and Shang E. Ha. Forthcoming. Mediation analysis is harder than it looks. In James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia (eds.), *Cambridge Handbook of Experimental Political Science.* New York: Cambridge University Press.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review* 102 (1): 33-48.

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

Jo, Booil. 2008. Causal inference in randomized experiments with mediational processes. *Psychological Methods* 13 (4): 314-36.

MacKinnon, David P. 2008. *Introduction to statistical mediation analysis*. New York: Lawrence Erlbaum.

Malhotra, Neil, and Jon A. Krosnick. 2007. Retrospective and prospective performance assessments during the 2004 election campaign: Tests of mediation and news media priming. *Political Behavior* 29:249-78.

Muller, Dominique, Charles M. Judd, and Vincent Y. Yzerbyt. 2005. When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology* 89 (6): 852-63.

Sobel, Michael E. 2008. Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* 33:230-51.