Article

Hammill Institute on Disabilities

Behavioral Disorders
2018, Vol. 43(4) 457–465

DOI: 10.1177/0198742917749560
journals.sagepub.com/home/bhd
SAGE

Causal Mediation in Educational Intervention Studies

Greg Roberts, PhD[1], Nancy Scammacca, PhD[1], and Garrett J. Roberts, PhD[2]

Abstract

Understanding the factors that mediate the effect of educational or behavioral intervention is critical to advancing both research and practice. When properly implemented, mediators add depth to the results of intervention research, indicating why a program works, highlighting ways to enhance its effectiveness, and revealing the elements that are essential to successful implementation. However, many researchers find mediation a difficult topic and struggle to implement it properly in statistical models of effects from between-groups randomized studies. In an effort to bring clarity to the topic of mediation and encourage its use where appropriate, this article lays out the requirements for evidence of a causal-mediated effect. An example of a randomized trial of an intervention targeting self-regulation and student behavior is used to illustrate the process of conceptualizing and testing for mediation of treatment effects. Statistical considerations also are addressed.

Keywords

statistical mediation, experimental design, quantitative methods

Funders, policy makers, and other stakeholders concerned with the results of educational and behavioral research are increasingly interested in factors that mediate a treatment's effect on key outcomes. For researchers exploring such effects, mediation can cause a good deal of confusion and consternation. The difference between mediation and moderation, the interpretation of mediated effects, and the statistical modeling of mediators in tests of treatment effects are often sources of uncertainty. In the context of between-groups intervention research (vs. single-cases or single-group pre/post designs, for example), where students or groups of students are assigned randomly to one or more treatment conditions, a mediator explains all or part of the treatment's impact on an intended outcome. In contrast, a moderator is a factor that reflects who is most affected by the treatment. In simple terms, a mediator explains *how* or *why* an intervention works, whereas a moderator explains *who* the intervention benefits or *what* conditions must exist for the intervention to be effective (Kraemer, Kiernan, Essex, & Kupfer, 2008). In addition, a moderator typically is a factor that exists prior to the introduction of an intervention, whereas a mediator, in this context, is an intermediate outcome that is measured or observed after the onset of the intervention. Treatment changes the mediator; changes in the mediator influence changes in the outcome.

In experimental designs (i.e., randomized groups), mediation is causal because randomizing provides a reliable counterfactual for identifying the model and for making inferences about the treatment's role in causing changes in both the mediator and the outcome (see Pearl, 2012, for a discussion of causal mediation in the general context of potential outcomes). One example of a mediator is fidelity of implementation. Fidelity can be viewed as representing the dosage of the treatment provided in the intervention and business-as-usual (BaU; that is, control) groups (Roberts, Lewis, Fall, & Vaughn, 2017). In the intervention group, it represents the extent to which the implementer adhered to the intervention protocol; in the BaU group, it represents the extent to which key elements of the treatment occurred due to intentional or accidental crossover. Fidelity fits the definition of a mediator because it is one mechanism through which the treatment produces the outcome. It is measured 1 or more times after the treatment begins and before the posttest is administered. Examples of moderators include student characteristics, such as special education status. In studying a treatment provided in general education classrooms, researchers

[1]The University of Texas at Austin, USA
[2]University of Denver, CO, USA

Corresponding Author:
Greg Roberts, The Meadows Center for Preventing Educational Risk, The University of Texas at Austin, 1 University Station D4900, Austin, TX 78712, USA.
Email: gregroberts@austin.utexas.edu

might investigate special education status as a moderator to determine if the treatment had a differential effect on the students with disabilities who were part of the study sample.

In this article, we describe the basics of causal mediation, with a focus on its application in school-based experimental studies of behavior-related interventions. We discuss the steps necessary for demonstrating a mediated effect in a dataset that typifies school contexts (e.g., multilevel, multicohort) using a hypothetical program of intervention research on self-regulation and student behavior to illustrate key points. Finally, we "put it all together" by describing several possible scenarios and discussing what each might suggest to the educational researcher.

The Basics of Causal Mediation

Causal mediation generally involves three variables: an independent variable, a dependent variable, and the mediator (models with multiple mediators can be fit as well; Preacher & Hayes, 2008). In school-based intervention studies (throughout this article, we use "intervention studies" as a proxy for between-groups, randomized designs), the independent variable is often an existing or newly developed treatment or program. A subset of participants is randomly assigned to receive the new intervention and a nonoverlapping group of similar students is assigned to a different condition, typical practice (BaU) when working in school settings. The dependent variable is an outcome that we expect the treatment to impact. In school settings, most achievement outcomes and some behavioral outcomes are continuous variables because they are scores from standardized assessments. Continuous variables usually distribute normally around the sample's mean according to its variance, a feature that can simplify analysis and lend greater statistical power. However, other data types are possible, including count data, ordered categorical (high, medium, low) variables, or binary (present/not present) outcomes. Behavior-related outcomes, in particular, may be noncontinuous because (a) behaviors can be counted; (b) severity, as a construct, is often scaled in categorical terms (e.g., extremely severe, severe); or (c) the outcome may be the absence of a once-present behavior (or the presence of a once-absent behavior).

The mediator, like the outcome, must be a malleable factor that treatment changes. By randomizing participants to conditions, we can assume that the treatment "causes" observed changes in the mediator (see Shadish, Cook, & Campbell, 2002). Also, like the outcome, the mediator can be continuous, ordered categorical, or nominal. Furthermore, as suggested earlier, we assume that changes in the mediator occur prior to changes in the outcome and that observed changes in the outcome are related, at least in part, to changes in the mediator. In this sense, the mediator explains one way by which a treatment may cause changes in an outcome. Assumptions about change in the mediator and its relationship to change in the outcome come with several important caveats that we discuss in more detail later in the article. We also save for a later section of the article a discussion of the measurement challenges related to the types of mediators often of interest to educational researchers.

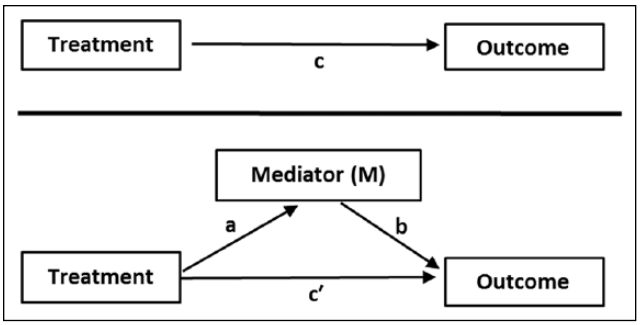

Figure 1. Causal mediation in randomized experimental design.

Figure 1 illustrates the key connections in a mediated relationship. Path *c* in the upper panel represents the *direct effect* of the treatment on the outcome, which is the difference between the average outcome for individuals participating in the intervention and the average outcome for the BaU group. Path *c* answers the question, "did the program work?" A statistically significant coefficient for Path *c* (in a regression model, for example) generally represents evidence that the program did work. It also suggests important, though often unasked, questions about "why" the program may have worked—questions that can be addressed using causal mediation analysis. In Figure 1, Paths *a* and *b* in the lower panel indicate the mediated effect (the way that the treatment affects the outcome via the mediator). The product of *a* and *b* (*a* multiplied by *b*) defines the *magnitude* of this effect. Because mediation involves two pathways, Path *a* and Path *b* through the mediator, it is described as an indirect effect. The *c′* path is the direct effect of the treatment on the outcome when the indirect effect is included in the model (it represents the "leftover variance" after accounting for the *ab* effect). Path *c* in the upper panel equals the total effect, which is merely the sum of the direct and indirect effects, $ab + c'$.

Mediation is sometimes discussed as *full* and *partial* mediation. When the indirect effect (*ab*) differs statistically from 0, suggesting the presence of causal mediation, *and* when *c′* does *not* differ statistically from 0, the indirect effect has completely displaced the statistically significant effect of *treatment* on *outcome* in Figure 1. This result can be described as full mediation because the mediator explains all of the treatment's effect. In practice, full mediation is uncommon. Partial mediation occurs when both the indirect effect, *ab*, and the direct effect, *c′*, differ statistically from 0. In this case, the mediator explains part, but not all, of the treatment's impact on the outcome, leaving a significant

amount of the treatment's variance in the direct effect. Direct, indirect, and total effects can be calculated for the entire sample and for different groups within the sample; effects can be compared across groups of interest. In intervention studies, the salient groups are often the groups assigned to different treatment conditions.

Modeling Causal Mediation

To illustrate, assume that we have developed an intervention to improve first-grade students' behavior in classroom settings during independent work times (e.g., small-group, student-centered activities). It involves direct instruction on classroom norms, instruction on a set of well-specified self-regulation strategies, along with structured, iterative practice on self-monitoring and self-correction. The program's logic is that greater awareness of behavioral expectations, improved self-monitoring, and an enhanced ability to self-correct as part of improved self-regulation will lead to fewer incidents of acting-out behavior during small-group, student-directed activities. A reasonable test of the new program's efficacy would be to contrast two groups of students who struggle to regulate their behavior, one group assigned to the new treatment and the other to BaU or to an alternative program that teaches self-regulation or another set of skills. The outcome would be improved behavior, which we might operationalize as fewer disruptive behaviors per targeted child for every hour observed during the week following the end of the new program's implementation.

The Direct Effect of Treatment

For purposes of this example, we assume (a) we have a sample that is adequately powered to detect the effect that we expect to demonstrate (we assume a sample of 100 students; see Kenny & Judd, 2014, for information on statistical power); (b) we have documented the fidelity with which the program was implemented with treatment-assigned students (or classes, schools, etc.); (c) the program was not implemented with students (classes or schools) in the BaU; and (d) we have reliably measured constructs. We also assume that we have baseline (i.e., prior to the onset of the intervention) counts of the behaviors in question for the sampled students and that the average number of disruptive behaviors prior to treatment (the baseline) does not differ for the treatment and comparison groups, suggesting that the randomization of students to treatment conditions yielded nondifferent groups, at least in terms of the outcome of primary interest. "Baseline" for count data is analogous to "pretest" when working with continuous data. The important consideration is that the two groups do not differ statistically prior to treatment on the targeted outcome.

Under these conditions, and assuming low and/or nondifferential attrition (What Works Clearinghouse, 2013, 2014) and no significant clustering (see Hedges, 2007), we can estimate an unbiased direct effect for the new treatment (Path *c* in the upper panel of Figure 1). In most cases, this would be a regression coefficient, and we assume that the coefficient is derived using ordinary least squares. (An understanding of particular statistical methods is not necessary for a basic understanding of causal mediation. Whether using ordinary least squares, logistic regression, multilevel modeling, or structural equal modeling, the steps for testing mediation are the same.) In our example, we would regress the outcome on treatment condition, treating the data as continuous. To simplify this discussion of the analysis and its results, we assume that the behavioral counts represent points along an underlying continuous distribution that approximates normal and, therefore, can be analyzed using ordinary least squares regression (see Sturman, 1999, for a discussion of the assumptions related to treating count data as continuous).

Suppose we discover that the 50 students who participated in the new program averaged 20 disruptive episodes per observed hour (for a total of 1,000 observed behaviors per hour observed across 50 sampled pupils) during the week just after the intervention's end. Meanwhile, the 50 students in the comparison condition average 40 disruptive behaviors per hour during the same week (we assume that the measurement protocols for documenting students' behavior are reliable across raters). Suppose as well that the baseline for both groups was 40. This means that the treatment group "improved" by an average of about 20 behaviors (we interpret a decrease on inappropriate behavior as improved behavior), whereas there were no changes, on average, in BaU group. The effect in raw units (number of behaviors) is 20.

Per the recommendations of the American Psychological Association (APA) Publications and Communications Board Working Group on Journal Article Reporting Standards (2008) and the What Works Clearinghouse (Institute of Education Sciences, 2014), we would also calculate and report an effect size, either a standardized group mean difference like Cohen's *d* or Hedges's *g*, the proportion of total variance explained by the independent variable (such as η^2 or η_p^2), or perhaps an odds ratio or relative risk statistic. The standardized mean difference, which can be expressed as a Cohen's *d* or Hedges's *g*, is 20 divided by an estimate of the sample variance. For Hedges's *g*, this estimate is a pooled estimate of the treatment and comparison samples' variance corrected for small-sample bias using a gamma function (Hedges, 1981). There are online calculators that calculate Hedges's *g* using the group means, standard deviations, and sample sizes as input.

Unlike tests of statistical significance, mean-based effect sizes are less dependent on sample size. For example, an effect of 0.65 (which suggests a pooled standard deviation of about 30 in our example) may be statistically significant

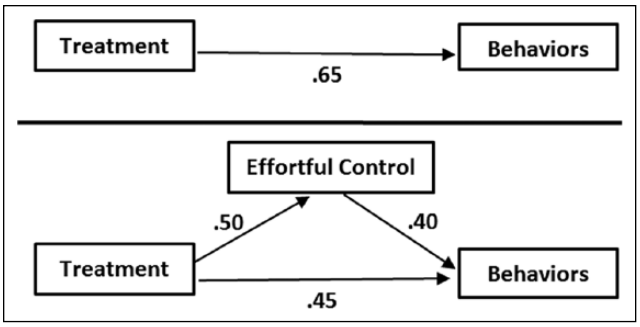

Figure 2. Running example with self-regulation as a mediator.

in a sample of 100, depending on several considerations such as how students were allocated to conditions and properties of the study's measures, whereas the same effect size may not differ significantly from 0 in a sample of 50. The effect is the same (a difference of about 65% of one sample-based standard deviation), but our conclusion regarding statistical significance would differ. In the former, we would describe our findings as evidence of the treatment's efficacy—the treatment caused improved behavior in our example. In the latter case, where a medium- to large-size effect did not differ statistically from 0, we would likely argue that the treatment is promising, though subject to additional study with properly powered research designs.

The Indirect Effect of Treatment

The top panel of Figure 2 summarizes the running example thus far. We have determined that the standardized mean difference for the direct effect is 0.65. The *unstandardized* regression coefficient would be in the neighborhood of 20, although the actual value would depend on the influence of other measured and modeled factors and on the estimation of an intercept (about 10 in this case because 10 is the average outcome when all other factors in the equation are 0). Raw units are useful when considering a single pathway (e.g., Path *a or* Path *b*), but when combining coefficients or when estimating a model with variables that are measured in different metrics, standardized coefficients ("beta coefficients") are more interpretable (see Preacher & Kelley, 2011, for discussion of the different types of standardized coefficients). Also, note that a standardized mean difference for two groups (i.e., an effect size) and the standardized regression coefficient for Path *c* in Figure 1 will often *not* be the same value (though they should represent the same "quantity"). However, when contrasting two groups of equal size from a person-randomized design, the regression coefficient can be estimated as a standardized mean difference between the two groups, and we assume as much in Figure 2.

A statistically significant direct effect suggests that the two groups differ at posttest. If the groups were nondifferent in all respects other than their exposure to the new treatment, both prior to the onset of treatment as well as during the treatment interval, then the new program is the most compelling candidate for explaining differences in students' behavior. "The program caused improved behavior" would be a reasonable conclusion. Left unanswered, however, are questions about *how* the program created that difference. For example, "what were the mechanisms by which the program had an effect?" It is this question that causal mediation can help to address. In the next sections, we consider Path *a*, Path *b*, Path *ab*, and Path *c′* in terms of our running example. We also "put it all together" by describing alternative outcomes for our hypothetical study and how each outcome might be best interpreted.

Path a. A first task in mediation analysis is to identify the set of potential or likely mediators. Recall that our program logic described activities as creating "greater awareness of behavioral expectations, improved self-monitoring, and an enhanced ability to self-correct as part of improved self-regulation." There are three potential mediators in this series: (a) awareness of behavioral expectations, (b) improved self-monitoring, and (c) enhanced ability to self-correct and self-regulate. Each appears to meet many of the requirements for a mediator—generally malleable, theoretically malleable in the context of the treatment's components, and temporally antecedent to the outcome. Ideally, we might measure and model all three as part of our study. For example, we could test students' knowledge of classroom expectations in both the BaU and treatment conditions. Because there are "correct" and "incorrect" responses to questions about classroom expectations (in the treatment condition at least), fairly straightforward measurement methods such as multiple choice spoken-word surveys (questions and response options are read to respondents) could be used, and their reliability easily established.

Self-monitoring and the meta-cognitive processes that support self-monitoring have been difficult to measure (Dent & Koenka, 2016). Tools in this area often rely on self-report; thus, they have high or unknown measurement error due to poor reporter reliability, limited construct validity, or social desirability bias (Dent & Koenka, 2016; Nisbett & Wilson, 1977; Veenman, 2011). Whereas measures of self-monitoring for adults and older aged schoolchildren are available and have been used with apparent success, it is unclear if young children are cognitively able to accurately report on their strategy use (Dent & Koenka, 2016). The argument is not that first graders do not self-monitor or that they cannot improve their capacity to self-monitor (Dignath, Buettner, & Langfeldt, 2008). Instead, the problem is measuring the self-monitoring of 6- and 7-year-olds with the reliability necessary to estimate its mediating effect.

Like self-monitoring, self-regulation can be difficult to measure (Greenberg, Kusché, & Speltz, 1991; Webster-Stratton & Taylor, 2001). However, because self-regulation

has been the subject of considerable research among preschool-aged and primary-grade students, and because much of the recent research has addressed measurement-related questions (e.g., Blair, Zelazo, & Greenberg, 2005; Cole, Martin, & Dennis, 2004; Raver, Jones, Li-Grining, Zhai, & Pressler, 2011; Wiebe, Espy, & Charak, 2008), we would be on more solid footing compared with self-monitoring (we recognize that self-monitoring is often included in models of self-regulation models, but we exclude it for purposes of this illustration). Supposing that our hypothetical intervention was developed in the context of the research on executive functioning in young children (Blair et al., 2005; Cole et al., 2004; Raver et al., 2011; Wiebe et al., 2008), we might select the Preschool Self-Regulation Assessment (PSRA; Smith-Donald, Raver, Hayes, & Richardson, 2007) to measure self-regulation. The PSRA yields scores for components of self-regulation, including *effortful control, attention/impulsivity*, and *executive functioning*. Because effortful control appears to be malleable (Somech & Elizur, 2012), we might devote a large component of our hypothetical treatment to improving students' effortful control and our first research question might be about the treatment's impact on effortful control: "Does the treatment cause changes in effortful control?" This effect is Path *a* in Figure 1. As before, with Path *c*, the analysis would likely involve regression. In this case, we would regress scores from the PSRA subtest for effortful control on students' assignment to treatment.

Recall that a mediator is defined as malleable in the presence of treatment and we assume that the mediator causes the outcome to some degree. This means that the mediator has to be measured at some point or at points after the onset of treatment in both the treatment and control groups. An additional measurement at pretest may be useful for checking on the success of randomly allocating students to groups and for controlling pretest differences in the mediator if randomization was not successful. There may also be conceptual or measurement-related reasons to model *change in* the mediator versus status at points subsequent to Time 1, in which case multiple measurement occasions would be necessary.

Depending on the mediator, the exact timing of data collection may be important (MacKinnon, 1994). For effortful control, the minimum would be measurement at a point close to posttest. Pretest-only measures of effortful control would be inadequate for the reasons described above, although additional measures at pretest and at a point (or points) during the intervention might be useful. Having three or more measurement occasions for the mediator (e.g., pretest, midpoint, posttest) would allow us to model effortful control as a trend over time, capturing status at pretest, status at posttest, and rate of change (and with a fourth data point, the shape) from start to finish. Trend estimates provide more information and also tend to be more reliable estimates of the construct being measured. Of course, three administrations of the PSRA may not be feasible, because the measure is lengthy, because the test is not designed for multiple administrations over relatively short time periods, or due to other practical, conceptual, or measurement-specific reasons. These are decisions for the investigator and for leaders in the school where the research is being conducted. Measurement at midpoint only is often acceptable and sometimes desirable, but might not adequately represent change in the mediator to the extent that more treatment translates into different levels or patterns of effortful control.

In our example, we assume that effortful control data were collected on one occasion, at a point prior to collecting data on the posttest outcomes. This requires us to "trust" the equalizing effects of randomization and to assume that the treatment groups did not differ, on average, at pretest, on levels of effortful control. For now, assume that the treatment's unstandardized effect on effort control was 5—that students in the treatment group scored 5 standard score points higher on the PSRA at posttest than the BaU students, on average, assuming no differences at pretest. We would calculate a standardized mean difference, or effect size, for the effect on effortful control, as described earlier, as well as a confidence interval around that effect. If the pooled standard deviation was 10 in the raw score units, the effect size would be about 0.50. Also, as with Path *c*, assume that we estimate the model such that the standardized regression coefficient for Path *a* is the same value as the standardized difference in group means, or 0.50. The confidence interval for an effect of 0.50 in a balanced sample of 100 would be [0.102, 0.898]. Because 0 is not included in this interval, we can assume that 0.50 differs statistically from 0 or that 0.50 is statistically significant. Recall that causal mediation assumes that the treatment operates, at least in part, through the mediator by causing changes in M, the mediator. Accordingly, a statistically significant coefficient for Path *a*, which we have, is required to move forward with our analysis. If the treatment did not cause changes in the mediator (i.e., if Path *a* did not differ from 0), then effortful control would not mediate the treatment's effect in this sample. In such a case, we might consider other potential mediators, such as attention/impulsivity or executive functioning, assuming that these PSRA data were collected, and assuming that these variables are theoretically viable as mediators of our treatment's effect.

Path b. Path *b* is *no* more difficult to calculate than Path *a* or Path *c*. Again, we rely on regression, in this case regressing the outcome on effortful control as measured by the PSRA. However, Path *b* is more controversial than Path *a*, due largely to the fact that we cannot randomize cases to levels of the mediator prior to measuring outcomes, which eliminates the logical basis on which we infer that treatment

causes changes in the mediator or that treatment causes changes in the outcomes. The mechanism that allows for causal inferences related to Path *a* is not available when estimating or interpreting Path *b*, at least in authentic classroom settings. However, many have argued that indirect effects from treatment to outcomes are not as discrete or mechanistic as our path models would suggest (see Kraemer et al., 2008). Instead, what we depict as a joint effect of two separate pathways may in fact be a response through which treatment influences both Path *a* and Path *b*. Muthén (2011) argued more generally for naïve perspective, suggesting that mediation would become unavailable to educational researchers working in schools if we require that strata of the mediator be formally manipulated. We split the difference and discuss Path *a* as a causal effect, Path *b* as a covariance, and Path *ab* as indirect causal effect, following Muthén (2011). The assumptions we make about Path *b* will influence how we interpret our findings. However, they do not change the underlying math necessary for estimating the model.

As before, Path *b* will be a regression coefficient that represents the covariation between effortful control and our measure of classroom behavior. We could estimate Path *b* in the total sample, with the treatment groups combined, but our interest would be in the group-specific values for *a* and for *b* because the *ab* product will be based on these group-specific values. At the same time, it is worth noting that improved effortful control may lead to improved classroom behavior, regardless the cause of the improvement in effortful control. There may be statistical reasons that the groups differ in Path *b*, such as restriction of range in the nontreated group, but when considered in the "population," the *b* parameter may be fairly constant, after adjusting for person- and context-level factors. However, because our design is randomized, we would expect that these external influences on effortful control would not differ across the two groups. For present purposes, we assume that the unstandardized coefficient for Path *b* is 0.60 in the total group, which means that for every increase of one unit of effort control (one standard score point), there is a corresponding decrease of just over half of a problem classroom behavior. The standardized coefficient is 0.40.

Path ab. The indirect effect is the product of the standardized coefficients for Paths *a* and *b*. In our example, the indirect effect would be 0.20 (0.50 for Path *a* times 0.40 for Path *b*). Most statistical packages will estimate the indirect term along with the regression model. Furthermore, modern programs (Mplus, SAS, Stata) will estimate *ab* and Paths *a, b*, and *c′* simultaneously, so the indirect effect is the value of $a \times b$ adjusted for other effects in the model, which is customary in multiple regression models. The statistical significance of *ab* is determined in one of several ways, depending on the software used. The simplest approach assumes that *ab* differs from 0 if *both* Path *a* and Path *b* differ from 0. Referred to as the joint test of significance, it is generally considered a preliminary test rather than a confirmatory test of the null hypothesis that $ab = 0$ (Fritz & MacKinnon, 2007).

Bootstrapping (Shrout & Bolger, 2002) is a more reliable approach, and also allows for the estimation of confidence intervals (Hayes & Scharkow, 2013). Bootstrapping uses resampling with replacement across a large draw of samples (1,000 to 5,000 is common) to create a sampling distribution for the indirect effect. The mean of the bootstrapped distribution will not always equal the indirect effect, because the product of two normally distributed variables (assuming *a* and *b* are normal) does not always yield a normal distribution. As a result, the mean of the bootstrapped distribution and *ab* are not always equal, which introduces bias. However, a simple correction can be applied (see Hayes, 2014, for more details), and a confidence interval and a standard error can be estimated, leading to a *p* value for *ab*. Fortunately, bootstrapping is *not* something done by hand. Modern software packages include this feature, and best practice appears to favor its use over several alternatives (Hayes & Scharkow, 2013). Preacher and Hayes (2004) provided downloadable SPSS and SAS macros to test indirect effects (http://quantpsy.org). Also, Mplus and Amos use bootstrapped samples to evaluate indirect effects.

Measurement Error in the Mediator

When working with observed variables (vs. latent variables, assignment variables, etc.), measurement error is a reality, and the task for the educational researcher is to minimize its impact rather than eliminate error altogether. Perfectly reliable measures may not be necessary, but measures with strong reliability are required, particularly for measuring the mediator. In models with a poorly measured mediator (i.e., low reliability), Paths *b* and *c′* will be biased, where the effect of the mediator on the outcome (Path *b*) is underestimated and the effect of treatment on the outcome (Path *c′*) is over-estimated, assuming *ab* is positive.

These problems can be addressed in several ways. The most obvious is to use measures with well-documented reliability. However, many mediators are conceptualized as internal processes (e.g., executive functions), and self-report is often the most efficient means of collecting such data. Even though self-report measures are often reliable according to the usual indices, they are subject to response bias and may not always be sufficiently valid, particularly when young children are asked to report on relatively sophisticated processes. Collecting data on the mediator using more than one measure is a good idea for several reasons. First, data from different measures of the same construct can be cross-referenced to determine if the measures lead to similar rankings of respondents. For example, if a

sample of students is ranked from highest to lowest on their levels of effortful control, these rankings should be similar across each measure to the extent that each is measuring effortful control. Multiple measures of a construct may also make it possible to create a latent variable for the mediator. Latent variables, which are featured in structural equation models, allow the educational researcher to specifically model measurement error, essentially eliminating error from the measurement of the mediator. Although more sophisticated than multiple regression, structural equation modeling (SEM) is a powerful tool for estimating causal mediation models. As indicated earlier, the basic steps do not differ from those outlined above (though several additional steps will be required by SEM), and the advantages of SEM often warrant the additional effort and cost.

The problem of omitted variables is another measurement issue that is particularly salient in a causal mediation context. This situation occurs when an unmeasured (or omitted) variable causes both the mediator and the outcome. When the independent variable is randomized, as in our example, omitted variables do *not* bias estimates of Paths *a* and *c*. That is an advantage of randomized designs. However, Paths *b* and *c′* can be biased if there is an omitted variable that causes both the mediator and the outcome, where Path *b* is overestimated and Path *c′* is underestimated. An omitted variable would also mask full mediation even in cases where the true value (in the population) for Path *c′* is 0. This problem is a key reason that statistical modeling should be based on theory and informed by prior research. Causal mediation is not a purely empirical exercise. It is a tool for evaluating the extent to which a conceptual model grounded in theory is supported by sample data. In our example intervention study described above, we noted that the PSRA provides scores for three components of self-regulation, but we modeled only the subtest score for effortful control as a mediator. Subtest scores for attention/impulsivity and executive functioning were omitted. If researchers had theoretical support for the intervention's effect on these variables, they could be included in the model as potential mediators to see if omitting them may have biased the results. Sensitivity analysis can also be used to evaluate the potential that a model has omitted variables and to estimate their effect (Muthén, 2011).

Putting It All Together

We have talked about the importance of evaluating individual pathways as a preliminary step in modeling causal mediation. However, mediation is a model-level phenomenon, and unbiased estimates depend on fitting a properly specified model. Our running example is essentially a path model, which we would fit as a multiple regression. It would be important to estimate the individual paths as a means of establishing the requirements for mediation (that Path *a* and Path *b* differ from 0, etc.), but to fit the model, we would regress the behavioral outcome on effortful control and on treatment (i.e., on the "upstream" variables), regress effortful control on treatment, and specify the indirect effect, *ab*, as a model parameter. Also, because we are dealing with two independent groups (treatment and BaU), we would fit the model in multiple groups, where pathways (Paths *b, ab*, and *c′* at least) are estimated in both groups (this model could be alternatively described as a moderated mediation model where assignment to treatment moderates the mediating effect of effortful control). This approach differs from using dummy-coded variables to indicate group membership.

In our example, Path *a* is 0.50, Path *b* is 0.40, and the indirect effect of treatment via effortful control (*ab*) is 0.20 (see Figure 2). This means that Path *c′* is 0.45 ($c' = c - ab$), where Path *c′* is the direct effect when the indirect effect is modeled, and Path *c* is the total effect. If 0.20 and 0.45 are significantly greater than 0 (they would be), we have partial mediation through effortful control, which would be a reportable finding. Our new treatment effectively decreases poor classroom behavior in part by increasing effortful control. Knowing this finding can help in several ways. First, if effortful control is a "lever" for improving behavior, we might consider intensifying program components that target effortful control to achieve an even greater total effect. If the other PSRA variables contribute to the mediating effect of our treatment, we may find that we can account for more of the treatment effect via the PSRA. This does not mean that the treatment is as effective as it could be or even should be. However, it does indicate that we have explained much if not all of the variance in the observed effect for the treatment as it is currently configured. If the other PSRA variables do not contribute to the mediating effect of effortful control, we would be interested in replicating the study to include other potential mediators and greater statistical power, if possible. Finding that effortful control is a key mechanism may also suggest other research possibilities, to the extent that similar increases in effortful control can be achieved via other, perhaps less intensive or less costly, interventions.

Table 1 summarizes a number of possible solutions. They demonstrate the logic of causal mediation, meaning that we did not model real data nor did we simulate data to produce these results. The sets of values were calculated as $c = ab + c'$. Model 1 is our running example. Models 2 through 9 describe other possible scenarios, with Paths *a* and *b* varied to include large- to small-size effects and the indirect and direct effects associated with each. Models 2 and 3 are similar to the running example in the treatment's effect on effortful control (0.50). They differ in the value for Path *b*. In Model 2, effortful control is a strong predictor of improved behavior. Because we measured effortful control on only one occasion, we could not say with confidence that

Table 1. Hypothetical Values for $c = ab + c'$.

Model	Path *a* (total)	Path *b*	*ab*	Path *c'*
1	0.50	0.40	0.20	0.45
2	0.50	0.70	0.35	0.30
3	0.50	0.10	0.05	0.60
4	0.10	0.40	0.04	0.61
5	0.10	0.70	0.07	0.58
6	0.10	0.10	0.01	0.64
7	0.90	0.40	0.36	0.29
8	0.90	0.70	0.63	~0
9	0.90	0.10	0.09	0.56

Note. We assume a total effect (*c*) of 0.65 for all models.

within-child increases in effortful control correspond to improved behavior at posttest. Instead, we would describe the coefficient as an average effect—higher levels of effortful control at a point close to posttest correspond to better classroom behavior. As a result of measuring effortful control after the start of the study, we would *not* be able to identify poorly behaved children in the BaU who nonetheless began the study with high levels of effortful control. The smaller Path *b* coefficient (0.10) in Model 3 means a smaller indirect effect and a much greater amount of the total effect left unexplained (0.60).

Models 4 through 6 represent cases where the treatment effect on effortful control is small (0.10). Notice that a small-size Path *a* limits the possible range of indirect effects (0.01–0.07). We do not know if this group of *ab* values differs statistically from 0 because it was not derived from actual analyses so we are not able to create confidence intervals or estimate *p* values. However, it is clear that when the treatment has a minimal effect on the mediator, indirect effects are small, which makes sense. The idea behind causal mediation is that the mediator "carries" the treatment's effect on outcomes, a challenge when treatment's effect on the hypothetical mediator is small. By the same logic, large treatment effects (Models 7–9) on the mediator are associated with relatively large indirect effects, particularly when the relationship of the mediator and the outcome is moderately large. The mediator in this case operates as a vehicle for transmitting treatment's effect to the outcome. Note that a large-size Path *a*, in this scenario, is necessary for full mediation (Model 8).

To conclude, we return to an earlier point about distinguishing between a treatment effect's practical significance and modeling its statistical properties. In our example, the total effect was 0.65, which represents a relatively healthy effect in most educational settings. However, it would be up to us, as the investigators, to make the argument that an effect of 0.65 warrants the time and effort involved in implementing the new intervention. This task differs from the purpose of causal mediation, though the results from mediation models may help when making the more general argument. However, causal mediation is a tool for explaining an observed effect. It can be used to determine why a program works to whatever extent it does work, producing findings that can advance a program of research, identify areas for improving the treatment, or highlight aspects of the treatment that are particularly salient to its effect and essential to its successful implementation.

Authors' Note

The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, the National Institutes of Health, the Institute of Education Sciences, or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Grant P50 HD052117 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100013 to The University of Texas at Austin as part of the Reading for Understanding Research Initiative.

ORCID iD

Nancy Scammacca https://orcid.org/0000-0002-7484-5976

References

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839–851. doi:10.1037/0003-066X.63.9.839

Blair, C., Zelazo, P. D., & Greenberg, M. T. (2005). The measurement of executive function in early childhood. *Developmental Neuropsychology*, *28*, 561–571. doi:10.1207/s15326942dn2802_1

Cole, P. M., Martin, S. E., & Dennis, T. A. (2004). Emotion regulation as a scientific construct: Methodological challenges and directions for child development research. *Child Development*, *75*, 317–333. doi:10.1111/j.1467-8624.2004.00673.x

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, *3*, 425–474. doi:10.1007/s10648-015-9320-8

Dignath, C., Buettner, G., & Langfeldt, H. (2008). How can primary school students learn self-regulated strategies most effectively? A meta-analysis on self-regulation training programmes.

Educational Research Review, *3*, 101–129. doi:10.1016/j.edurev.2008.02.003

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*, 233–239. doi:10.1111/j.1467-9280.2007.01882.x

Greenberg, M. T., Kusché, C. A., & Speltz, M. (1991). Emotional regulation, self-control, and psychopathology: The role of relationships in early childhood. In D. Cicchetti & S. L. Toth (Eds.), *Internalizing and externalizing expressions of dysfunction: Rochester symposium on developmental psychopathology* (Vol. 2, pp. 21–55). Hillsdale, NJ: Lawrence Erlbaum.

Hayes, A. F. (2014). The simple mediation model. In A. F. Hayes (Ed.), *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (pp. 85–122). New York, NY: Guilford Press.

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science*, *24*, 1918–1927. doi:10.1177/0956797613480187

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Education Statistics*, *6*, 107–128.

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, *32*, 151–179. doi:10.3102/1076998606298040

Institute of Education Sciences. (2014). *What Works Clearinghouse procedures and standards handbook* (Version 3.0). Washington, DC: U.S. Department of Education.

Kenny, D. A., & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychological Science*, *25*, 334–339. doi:10.1177/0956797613502676

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, *27*(2, Suppl.), S101–S108. doi:10.1037/0278-6133.27.2(Suppl.).S101

MacKinnon, D. P. (1994). Analysis of mediating variables in prevention and intervention research. *NIDA Research Monograph*, *139*, 127–127.

Muthén, B. (2011). *Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus*. Retrieved from http://www.statmodel.com/download/causalmediation.pdf

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259. doi:10.1037/0033-295X.84.3.231

Pearl, J. (2012). The causal mediation formula—A guide to the assessment of pathways and mechanisms. *Prevention Science*, *13*, 426–436.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, *36*, 717–731.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93–115. doi:10.1037/a0022658

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development*, *82*(1), 362–378. doi:10.1111/j.1467-8624.2010.01561.x

Roberts, G., Lewis, N. S., Fall, A. M., & Vaughn, S. (2017). Implementation fidelity: Examples from the reading for understanding initiative. In G. Roberts, S. Vaughn, S. N. Beretvas, & V. Wong (Eds.), *Treatment fidelity in studies of educational intervention* (pp. 61–79). New York, NY: Routledge.

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

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and non-experimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422–445. doi:10.1037//1082-989X.7.4.422

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly*, *22*, 173–187. doi:10.1016/j.ecresq.2007.01.002

Somech, L. Y., & Elizur, Y. (2012). Promoting self-regulation and cooperation in pre-kindergarten children with conduct problems: A randomized controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*, 412–422. doi:10.1016/j.jaac.2012.01.019

Sturman, M. (1999). Multiple approaches to analyzing court data in studies of individual differences: The propensity for Type 1 errors illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement*, *59*, 414–430. doi:10.1177/00131649921969956

Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instruments: A discussion. *Metacognition and Learning*, *6*, 205–211. doi:10.1007/s11409-011-9080-x

Webster-Stratton, C., & Taylor, T. (2001). Nipping early risk factors in the bud: Preventing substance abuse, delinquency, and violence in adolescence through interventions targeted at young children (0-8 years). *Prevention Science*, *2*, 165–192. doi:10.1023/A:1011510923900

What Works Clearinghouse. (2013). *Assessing attrition bias* (Version 2.1). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

What Works Clearinghouse. (2014). *Assessing attrition bias—Addendum* (Version 3.0). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology*, *44*, 679–692. doi:10.1111/j.1467-7687.2010.01012